How is fairness measured in practice? A survey and critique of fairness analyses applied to predictive algorithms

ANDREW PUTMAN, Ontario Tech University, Canada
PETER R. LEWIS, Ontario Tech University, Canada
DAVID RUDOLER, Ontario Tech University, Canada

The increasingly rapid pace of research and development of prediction algorithms over the past decade has spurred their adoption across seemingly every industry and aspect of life. This growth and the subsequent increase in public awareness have brought a range of concerns into focus regarding the potential bias and fairness impacts of these algorithms. This cross-disciplinary survey assesses peer-reviewed publications from the last 10 years which assessed fairness in a predictive algorithm using real world data. We begin with a descriptive overview of the fields of research, types of publication, and the types of outcomes being predicted. We then engage in a more detailed exploration of the fairness analyses that were applied, the metrics used, and the evidence provided to justify these selections. We also discuss the contexts in which these analyses were conducted. We conclude by highlighting the lack of inclusion of perspectives from people who are impacted by the algorithms being analyzed and discuss directions for future work in this area.

CCS Concepts: • **Computing methodologies** → **Philosophical/theoretical foundations of artificial intelligence**; • **General and reference** → *Surveys and overviews.*

Additional Key Words and Phrases: Fairness, Bias, Justice, Equity, Predictive Algorithm, Artificial Intelligence, Machine Learning, Measurement, Metric, Decision-Making

1 Introduction

The immense investment into prediction-based technologies over the past decade plus has resulted in several technical and implementation breakthroughs using predictive algorithms. These technologies may utilize predictive algorithms in complex, interconnected structures such as transformer neural-network architectures or may implement relatively simple, straight forward prediction techniques like linear regression [9, 30, 38]. Over the past few years in particular, many of the technologies that leverage prediction algorithms at some point in their process have been marketed to the public under the catch-all label of "AI" [22]. The increased prevalence of predictive algorithm-driven technologies is perhaps most obviously demonstrated by the mass uptake in use of large language model-based applications like Open AI's ChatGPT or Google's Gemini [23, 28]. The rapid, widespread increase in public (re)awareness of predictive algorithms that has been reignited by these applications has also increased the frequency and intensity of calls both to adopt these technologies more broadly and raised concern about the moral and ethical consequences of doing so.

In response to the surge of industry, academic, and public interest in the adoption and effects of predictive algorithms, many tools, frameworks, measures, and metrics have been developed to identify or address bias and fairness in predictive algorithms. Several of these tools have received institutional backing from major tech companies including Microsoft's Fairlearn [40], Amazon's SageMaker Clarify [18], LinkedIn's LiFT [37], and IBM's AI Fairness 360 [3] and Watson OpenScale [20] tools. These tools offer a vast array of methods and metrics that could be used in the assessment fairness in predictive algorithms. So many, in fact, that an increasing number of researchers, ethicists, and community leaders are becoming concerned that these types of tools more effectively enable "fairness washing" than they assist in meaningful

Authors' Contact Information: Andrew Putman, Ontario Tech University, Oshawa, Ontario, Canada, andrew.putman@ontariotechu.ca; Peter R. Lewis, Ontario Tech University, Oshawa, Ontario, Canada, peter.lewis@ontariotechu.ca; David Rudoler, Ontario Tech University, Oshawa, Ontario, Canada, david.rudoler@ontariotechu.ca.

investigation of fairness in predictive algorithms [5, 6, 27, 36, 41]. In this context, fairness washing refers to the *post hoc* cherry-picked reporting of only fairness measures that an algorithm performs well on, creating the illusion of having met a rigorous standard for algorithm prediction fairness without actually having done so.

Currently, existing surveys, guidelines, and assessment tools addressing fairness assessment in predictive algorithms provide a broad selection of possible fairness analysis methods and metrics to draw upon [3, 8, 24, 39, 40]. Despite or perhaps because of this breadth of choice, there does not appear to be any consensus as to which tools are appropriate for use in which contexts, nor is it currently clear how these analysis techniques are actually being used in practice. Therefore, we decided to formally investigate these questions by the applying a rapid review-style methodology to survey the predictive algorithm fairness testing literature.

1.1 Research Questions

This literature survey seeks to answer the following research questions.

(1) Which forms of fairness analysis are most prevalent in peer-reviewed literature that empirically assesses fairness in predictive algorithms using real-world data?
   (a) Which types of justifications were provided to support the selection of the fairness analysis methodologies implemented in the surveyed articles?
(2) Which metrics are most frequently used to measure aspects of fairness within the fairness analyses of peer-reviewed articles that empirically assess predictive algorithms using real-world data?
   (a) Which types of justifications were provided to support the context-specific validity of the fairness metrics used by these articles?

1.2 A Note Regarding Terminology

For the purposes of this paper we will draw upon the NIST definition of an algorithm [12], and use the term *predictive algorithm* to refer to any clearly specified mathematical process for computation; a set of rules that, if followed, will produce a prediction based on the information input into that process. A predictive algorithm's prediction must be reproducible. This means there needs to be a way to specify a seed or similar setting within the process such that the algorithm will always produce the same prediction given the same input and settings, regardless of whether this is how the process is typically interacted with.

For the sake of clarity throughout this paper we will exclusively use the term *predictive algorithm* to refer to any processes that meet the above definition.

2 Methods

This survey was conducted using an adapted rapid review (survey) methodology. Our survey methodology was developed using the Cochrane Rapid Reviews Methods Group guidelines [14, 15], with the largest difference between the guidelines and our methods being the splitting up of the article screening process into three rounds instead of two. This change to the screening process better reflects computer science writing norms as well as the ACM- and IEEE-styles of article formatting. The process of conducting the search, screening, and evaluation of articles is detailed in the following subsections.

Inclusion Criteria	Exclusion Criteria
Involves some form of prediction or decision-making. Evaluates predictions produced by an algorithm or algorithms. Provides an empirical definition of bias or fairness. Article was published in or after 2015. Article was published in English. Full text is available at Ontario Tech University.	Definition of bias or fairness was not applied empirically to real-world data. Definition of bias or fairness was not applied to an algorithm[s]' predictions.

Table 1. Inclusion and exclusion criteria used to create the survey's sample

2.1 Inclusion and Exclusion Criteria

The inclusion and exclusion criteria for this survey, which are presented in Table 1, were set in line with our research questions. The inclusion criteria were initially operationalized through the creation of the literature search strategy and were then applied throughout all stages of article screening (detailed in section 2.2). The exclusion criteria were applied in addition to the inclusion criteria during the final full text article screening stages to ensure that our final sample would reflect the population we wished to survey.

2.2 Search Strategy and Article Screening

The first iteration of the search strings and database selection for our survey were informed by our research questions, then this initial search strategy was further refined with the help of a subject-matter expert librarian at Ontario Tech University. We used our finalized search strategies to search the ACM Digital Library [34], Computer and Applied Sciences Complete [11], and SCOPUS [13] databases for peer-reviewed journal articles and published conference proceedings that were published between January 1st, 2015 and March 7th, 2025 (the date the searches were conducted). The articles retrieved by the search needed to have mentioned the terms related to bias, fairness, algorithm, and assessment in their title, abstract, or keywords. The complete search string for each database has been included in Appendix A.

The article screening was conducted in 3 stages to better accommodate the normative differences in abstract writing across academic disciplines. For example, in health sciences abstracts are typically highly condensed versions of the entire work, which specify the methodology used and the exact results of the experiment conducted. This contrasts with abstract writing norms across many disciplines of computer science, where abstracts are often written less as specific result-oriented summary and more as short introduction to explain what the article is about and why someone might wish to read it.

The first stage of article screening involved examining the potential applicability of the inclusion criteria broadly using the title, abstract, and keywords used by each article. The second stage of screening consisted of skimming the full-text of the article, looking for the presence of the inclusion criteria, and the final stage of screening was to read the entirety of each article and fully applying the survey's inclusion and exclusion criteria.

2.3 Data Extraction

Bibliographic information (title, authors, publication type, date of publication, etc.) for each of the articles included in the full survey was extracted from the metadata contained in the database that the article had found in.

2.3.1 Fairness Analysis Variables. The fairness analysis in each article surveyed was described qualitatively using the following criteria and their definitions.

- **Equivocates lack of bias with fairness.** *The article's definition of fairness for the purposes of its analysis assesses for a lack of bias between individuals or groups.*

For example, when assessing a binary classification algorithm an article may choose to define fairness as being when observations that have an attribute "A" have an equal probability of being predicted to be in the "positive" class by the algorithm as observations that do not have the attribute "A". (This definition is typically called group fairness or statistical parity difference [39]). Assessing the statistical parity difference metric in an algorithm's predictions represents a purely technical assessment of performance, in this case an assessment of prediction bias associated with the attribute "A". Therefore, unless the article grounds their assessment of statistical parity difference in the specific context of the prediction itself (the qualitative dimension that distinguishes fairness from bias), the example article has conflated a lack of bias with being fair in their analysis.

- **Fairness analysis methodological justification.** *A categorization of the degree to which each article has provided evidence that supports its chosen fairness analysis methodology.*

Articles were classified as "justified" if they (a) specified one or more specific ethical, philosophical, or theoretical frameworks to support its decision; (b) provided context-specific evidence of the validity of its chosen methodology; (c) consulted with persons affected by the algorithm to confirm their chosen method's appropriateness and applicability; (d) cited specific laws or legal standards that were being measured against; or (e) had cited previous fairness analyses for the purpose of conducting a replication study. Articles were classified as "partially justified" if they (a) provided evidence for the validity of methodology that was not specific to the context in which it was being applied; (b) had cited previous fairness analyses without explaining the connection between those analyses and the methodology selected ; or (c) claimed justification by alluding to ethical, philosophical, or theoretical frameworks without providing evidence to support this claim specifically. Articles were classified as having "no justification" if no justification for the fairness analysis methodology was provided in the text of the article.

- **Fairness analysis included those affected by the algorithm.** *The article indicated that the fairness analysis was planned with, conducted alongside, or had consulted with people who are impacted by the algorithm being analyzed.*

When this was the case, the persons who had been consulted were categorized by whether those people were persons who apply the algorithm, have the algorithm applied to them, or both. Articles that analyzed real-world data but were primarily focused on replication or meta-research into the research of fairness in predictive algorithms were excluded from this question.

- **Context of fairness analysis:** *A classification situating the context of that article's fairness analysis.*

A fairness analysis was classified as the "primary focus of the article" if the fairness analysis of the predictive algorithm is the purpose of the article. A fairness analysis was classified as being a "novel fairness-related intervention" if the article was proposing new predictive algorithms, measurements of fairness, or a novel framework for assessing fairness in predictive algorithms. A fairness analysis was classified as a "sensitivity analysis" if it was included as a secondary analysis or test of robustness that was not part of the primary research question. A fairness analysis was classified as a "meta-fairness analysis" if its primary focus was researching the ways in which fairness and predictive

algorithms are implemented. (e.g., a study comparing the performance of multiple fairness mitigation techniques using a real-world scenario/dataset).

2.3.2 Fairness Measure and Metric Variables. The specific measure and metrics used in each article's fairness analysis were identified and summarized by the average number of measures used per article and the frequency with which each measure appeared in the surveyed articles. In this survey out goal was to describe how fairness is being assessed in the literature, so rather than applying a specific taxonomy or categorization to the measures we observed, we have reported the measures as described by the article they were used in. The exception to this was if multiple names had been used to describe the same calculation, in which case the measures were combined under a single name for the purposes of this survey's analysis. The fairness metrics used in each article were also categorized qualitatively using the following criteria.

- **Fairness measure selection justification:** *A categorization of how each article supported its selection of the metric[s] used in its fairness analysis.*

An article's fairness metric selection was classified as having "evidence of context-specific validity" if it cited or provided evidence supporting the validity of the measures used. An article was classified as having "cited prior use" if the article cited that prior work had used the selected measures but did not provide evidence supporting the validity of those measures in the article's use case. An article was classified as having a "plausible argument" if it provided an explanation supporting its measure selection without providing evidence to support that argument. An article was classified as "none" if it did not provide an explanation for its measurement selection.

- **Level of detail provided for fairness measure:** *Indicates how an article describes the fairness measures used in its analysis.*

An article was classified as "formula" if the article provided the formula used to calculating the fairness measures employed in the article. An article was classified as "interpretation" if it included written guidance regarding how to interpret the fairness measure. An article was classified as having provided "both" if it met both the formula and interpretation criteria. An article was classified as "none" if no explanation of the fairness measure was provided.

3 Results

A step-by-step walk through of the literature search and screening process that produced this survey is presented visually in Figure 1 using an adapted PRISMA flow diagram [29]. The initial search of the three databases returned 3837 results, consisting of 3563 unique, non-retracted articles. After conducting the abstract screening, 631 articles were passed along for the full-text skim screening stage, of which 378 articles were selected for the full-text screening stage. Of those 378 articles, 35 articles were excluded for not including an empirical application and an additional 14 articles were excluded for not directly evaluating the predictions of an algorithm. In total, the screening process produced 329 articles which met the inclusion and exclusion criteria for this survey. A complete bibliography of the articles analyzed in this survey can be found in Appendix B.

As can be seen in Figure 2 (a), more than 90% of the articles included in this survey were published within the past five years (2021-2025). Further, nearly half of all the surveyed articles were published within 12 months of the search being conducted. Figure 2 (b) shows the adjusted frequency of publication for articles that assess predictive algorithms and fairness when controlled by the overall number of publications that assess predictive algorithms. More

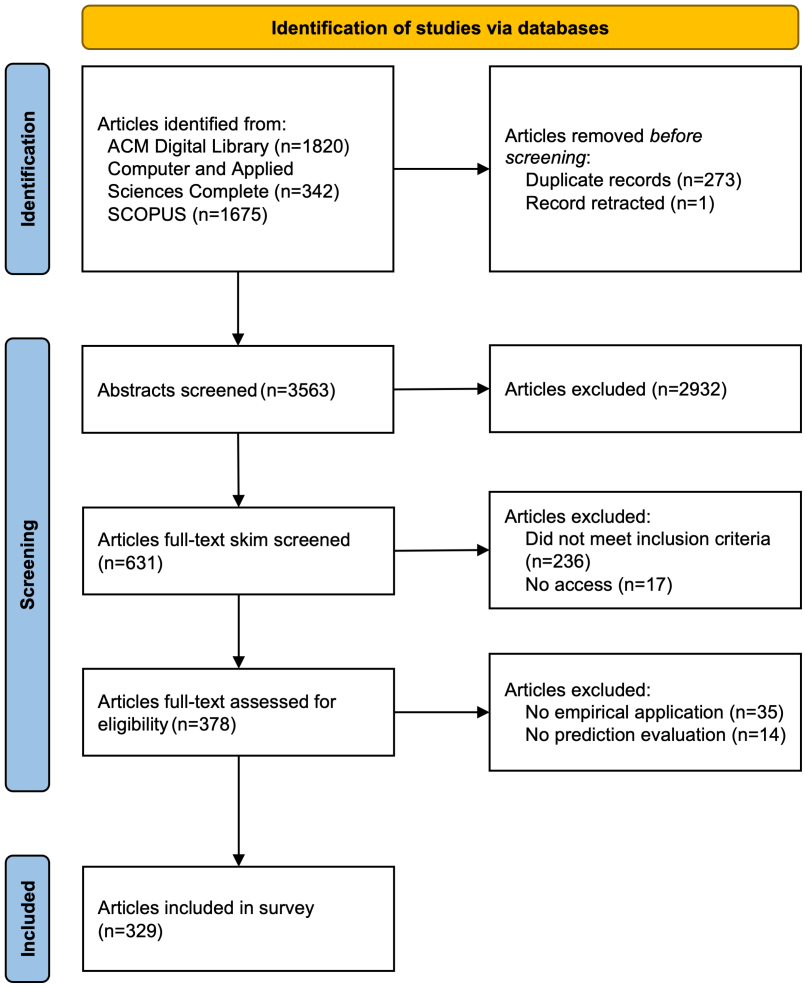

Fig. 1. Adapted PRISMA flow diagram of rapid review process. Diagram adapted from Page MJ, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021 Mar 29;372:n71.

specifically, that control was set by re-running the search strings used to create this survey's sample without any of the fairness-related search terms.

Slightly more than two thirds of all the articles included in the survey were journal publications. On a year-by-year basis a larger proportion of articles were published as part of conference proceedings until 2021, after which the number of journal articles outpaced the number of conference papers published per year in our survey. Articles in this survey came from a diverse set of disciplines including healthcare, criminal justice, education, human resource management,

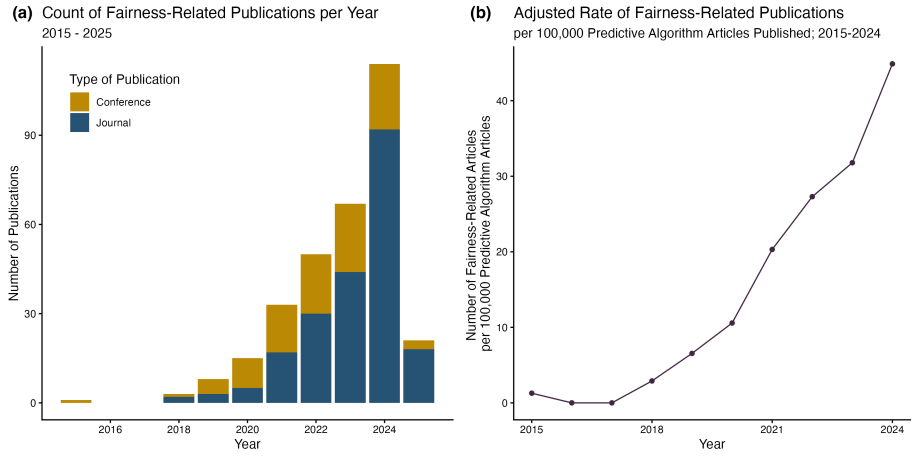

Fig. 2. Base frequency (a) and adjusted-rate (b) of peer-reviewed articles published that assessed fairness of predictive algorithms using real-world data. By year, 2015-2015.

Number of Articles	329
Fairness Analysis (%)	
Explicitly Incorporated	260 (79.0)
Hypothesized Applicability	52 (15.8)
Not Incorporated	17 (5.2)

Table 2. Inclusion of fairness analysis in surveyed articles

and social work; all of which intersect with computer science and related fields in the study of fairness in predictive algorithms.

As can be seen in Table 2, nearly four fifths (n = 260, 79.0%) of the articles in our survey explicitly incorporated some form of fairness analysis in their results. A further 52 articles (15.8%) hypothesized about the applicability of extending their work onto notions of fairness and there were 17 articles (5.2%) which met this survey's inclusion and exclusion criteria despite not actually reporting results from a fairness analysis. This is largely an artifact of the multitude of meanings of the word "bias" (see section 3.1 for more detailed discussion). For example, if an article discussed the importance of fairness, justice, or equity without evaluating any of those notions and the article's analysis involved evaluating the difference between an algorithm's predicted and actual means (what we'll refer to as "statistical bias"), then that article met this survey's inclusion and exclusion criteria through a technicality. Therefore, since this survey's goal is to understand how fairness analyses and measurements are being used in practice, these 17 articles were dropped from the rest of this survey's analyses. This change causes the reduction seen from our original survey sample (n = 329) to the revised analytic sample of (n = 312) articles.

Table 3 reflects that 90.1% of the articles in our literature survey assessed an algorithm's predictions of real-world data in the context a binary classification prediction task. Regression (4.2%), ordinal ranking (4.2%), and multinomial classification (2.2%) were the next most common prediction tasks tested. Ten (3.2%) of the articles in the survey reported testing fairness on more than one type of prediction task. The vast majority of articles (n = 275, 88.1%) measured the

Number of Articles	312
Fairness Analysis (%)	
Explicitly Incorporated	260 (83.3)
Hypothesized Applicability	52 (16.7)
Prediction Tasks (%)	
Binary Classification	281 (90.1)
Graph Matching	1 (0.3)
Image Segmentation	3 (1.0)
MAUP	1 (0.3)
Multinomial Classification	7 (2.2)
Ordinal Classification	1 (0.3)
Predicted Probability	3 (1.0)
Ranking	13 (4.2)
Regression	13 (4.2)
Type of Comparison Used (%)	
Algorithm – Algorithm	275 (88.1)
Human – Algorithm	8 (2.6)
No Comparison Used	29 (9.3)

Table 3. Descriptive summary of prediction tasks utilized by surveyed articles

fairness of an algorithm's predictions in relation to another algorithm or algorithms. Another 29 articles (9.3%) assessed prediction fairness without making a direct comparison to another prediction process and a small number of articles ($n = 8, 2.6\%$) evaluated a predictive algorithm's fairness in relation to human predictions for the same task.

3.0.1 Authors' Interpretation: Survey Sample. In this survey we observe a rapid growth in the number of publications investigating fairness in predictive algorithms that use real-world data. There were almost no articles that met our search criteria in 2015, whereas that number rose to more than 100 articles that met our search criteria in 2024. This trend of year-over-year increases in the number of publications continues to hold when we controlled for the overall rate of predictive algorithm papers published between 2015 and 2025. However, it is of note that to see the year-over-year trend visualized in Figure 2 (b), the proportion of publications assessing fairness had to be scaled to be the number fairness-related articles published per 100,000 total predictive algorithm articles published. This suggests that while there has been a real increase in research into fairness and predictive algorithms over the past five-plus years, its current 2024 peak of < 50 fairness-related articles per 100,000 total articles suggests that this is still a relatively niche area of predictive algorithm research.

As has been highlighted in previous surveys, the current body of work investigating fairness and predictive algorithms has been almost entirely focused on binary classification problems [2, 8, 24, 25]. This trend is reflected by our sample as well, as was expected. Our sample also demonstrates that there has been some work toward implementing fairness analyses more broadly to ordinal ranking and regression-based predictive algorithms, though these efforts typically seek to extend the binary classification-based notions of fairness onto these problem types [2, 25, 26, 35]. Further discussion regarding the fairness analyses and measurements employed in the surveyed articles can be found in Subsections 3.1.1 and 3.2.1, respectively.

We do not find it surprising that the vast majority of the articles in our survey assessed algorithm performance (and fairness) by contrasting two or more predictive algorithms. As we see it, this is mostly likely because (a) not all the real-world prediction tasks that were assessed are currently being done by humans (e.g., if a decision is currently being made by an algorithm, comparing an intervention against that current algorithm would be best practice), (b) comparing an algorithm's predictions to (expert) human predictions may be infeasible or beyond the scope of the research project being reported on in the article, and (c) the use of human predictions typically requires receiving research ethics board approval which can add considerably to the length of time and amount of preparatory work needed prior to conducting the article's analysis. When we further add the context of how relatively new this area of research is, we believe it makes sense that relatively few large scale studies have been conducted to do date where an aspect of fairness is assessed by comparing algorithm and (expert) human predictions.

3.1 Fairness Analyses

A descriptive summary of the fairness analyses that were reported by the articles in the survey can be found in Table 4. More than four fifths of the articles (254 articles, 81.4%) conflated a lack of bias with fairness. Just under half of the analyses (152 articles, 48.5%) presented a novel fairness-related intervention, and approximately one third of fairness analyses reported (105 articles, 33.7%) were the primary objective of the article they were reported in.

Nearly all the articles surveyed (305 articles, 97.8%) provided some form of justification for their methodological choice of fairness analysis. Of the 118 articles (37.8%) that provided a complete justification for their methods' selection, the most common type of justification was context-specific validity (48 articles, 15.4%), followed by having specified a particular ethical, philosophical, or theoretical framework (43 articles, 13.8%). Having cited a specific policy or law to be tested and conducting replication research were both employed in 14 (4.5% each) articles' justifications. The majority of the articles (187, 59.9%) in our survey partially justified their selected fairness analysis methodology, 155 of which cited the prior use of the fairness analysis methodology but omitted explanation as to why that fairness analysis methodology was applicable to the context in which the article was using it.

Less than one percent of the surveyed articles (2, 0.6%) included persons affected by the real-world algorithm predictions in their fairness analysis process. The two articles that included people affected by the algorithm being assessed in their evaluations both included persons who apply the predictive algorithm being assessed. None of the surveyed articles included people who had had the algorithm applied to them in their fairness analysis processes.

3.1.1 Authors's Interpretation: Fairness Analyses. One of the central themes we identified throughout this survey is the pervasiveness of narrowly defining fairness as a lack of bias. More specifically, this assertion posits that when an algorithm's predictions lack inter-group or between-individual differences for a specified statistical bias measure, then that algorithm can be said to be "fair" (at least in terms of the measure being tested). The ubiquity of this conflation is perhaps most poignantly demonstrated by the number of articles in this survey that acknowledge the existence of a wide variety of ways in which fairness could be defined and operationalized, then go on to solely assess fairness as a lack of bias (equivalence) between groups. Defining fairness as a form of equivalency between groups or individuals is not necessarily an invalid operationalization of fairness, though it is important to recognize that this definition relies on the *assumption that the aspect of fairness being measured can be meaningfully reduced into a single equality statement.*

By our assessment, the defining of fairness to be a lack of bias has most likely been arrived at using egalitarian perspectives on fairness. Broadly, egalitarian notions of fairness aim for groups and individuals to receive equal amounts of success or for those persons to be provided equal access (opportunity) to achieve said success [1, 4]. These notions

Number of Articles	312
Fairness Analysis (%)	
Explicitly incorporated	260 (83.3)
Hypothesized applicability	52 (16.7)
Conflated [Lacking] Bias with Fairness (%)	
Yes	254 (81.4)
No	58 (18.6)
Context of Fairness Analysis (%)	
Fairness meta-research	40 (12.8)
Novel fairness-related intervention	152 (48.7)
Primary focus of Article	105 (33.7)
Sensitivity analysis	15 (4.8)
Fairness Analysis Justification (%)	
Methodology justified	118 (37.8)
Context-specific validity	48 (15.4)
Framework	43 (13.8)
Law/Policy	14 (4.5)
Replication	14 (4.5)
Methodology partially justified	187 (59.9)
Claim without evidence	9 (2.9)
Non-specific validity	22 (7.1)
Previous analysis	155 (49.7)
No justification provided	7 (2.2)
Impacted Persons Included in Analysis (%)	
Yes	2 (0.6)
People who apply the algorithm	2 (0.6)
People who the algorithm is applied to	0 (0.0)
No	310 (99.4)

Table 4. Descriptive summary of fairness analyses

map roughly onto the rights-based conceptualizations of fairness and meritocracy that are prevalent throughout North American and European law and social norms [1, 4, 31–33]. Given the high proportion of the authors in our survey whose reported institutional affiliation were from North America or Europe, it makes some sense that an egalitarian approach to fairness could be the default perspective for many of these researchers.

Several prominent surveys have touched upon the question of how researchers who were investigating fairness in predictive algorithms/ML/AI have operationalized fairness within their analyses. These include surveys by Verma & Rubin [39], Mehrabi et al. [24], and Caton & Haas [8], all of whom report evidence in line with what we found in this survey; that the majority of research investigating fairness in predictive algorithms narrowly defines fairness as being an equality between groups or individuals (a lack of bias). Mehrabi et al. goes further by asserting that "[i]n the context of decision-making, fairness is the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics" [24, pg.1]. Notably, there is no citation provided for this definition, nor is there explanation as to the theoretical framework or evidence base that justifies why the provided definition can be said to be "the" definition of fairness. We use Mehrabi et al.'s definition here not as a critique of that work in

particular, rather we believe their definition is representative of the pervasive conflation between (a lack of) bias and fairness throughout this area of research [24]. Definitions steeped in an egalitarian, equivalency-based perspective of the meaning of "fair" represent a lens through which some aspects of fairness can be conceptualized, however, we question whether *the ubiquity of these equivalency-based operationalizations of fairness is reflective of those definitions being the most appropriate and valid choice for the context in which they have been used.*

When we consider the immense variety of discourses on the topic of fairness across time and culture, it is perplexing to see how little the surveyed articles draw from this rich history. This lack of engagement from the surveyed articles can be seen in how more than half of the articles did not make space in their manuscript to discuss why their selected operationalization of fairness was appropriate for the context in which it was being applied. Many of these articles take the time to (correctly) emphasize that there is no one single definition or operationalization of fairness that can be comprehensive of every situation, unfortunately, when these articles also neglect to provide specific justification for the fairness operationalization they use in their analysis, they seem to imply that 'if there is no one, universal definition of fairness then all fairness definitions must be equally valid'. We posit that on the contrary, *a lack of a single, universally generalizable definition of fairness makes it even more crucial that we deeply consider the context-specific validity of the fairness analysis methods we employ*. The seemingly handwave-y dismissal of the responsibility to justify the validity of the fairness analyses used becomes even more troubling when said article goes on to make factual claims about having "improved" prediction fairness.

Perhaps the most surprising observation from this survey is that only 2 of the surveyed articles mentioned incorporating the perspectives of people were or would be involved with the use of the algorithm being studied, and 0 of the articles included the perspectives of people who were or would be affected by the assessed algorithm's predictions. While the incorporation of those impacted by or who could have been impacted by a predictive algorithm may not have been feasible for every fairness analysis within our survey, it seems surprising that less than one percent of the analyses reported attempting to do so. This lack of incorporation of lived experiences seems particularly surprising when we take into account the simultaneous general lack of theoretical justification provided to support the fairness analyses being used.

3.2 Fairness Measures and Metrics

As can be seen in Table 5, three quarters of the 312 surveyed articles (235 articles, 75.3%) justified the metric(s) used in their fairness analyses through citing the prior use of that metric without providing evidence that supported the metrics use in the context in which the article was applying it. Less than 10% of the articles surveyed provided direct evidence supporting the validity of the metrics they employed or were a conducting a replication study (15 articles, 4.8% and 11 articles, 3.5%, respectively). Nearly all the surveyed articles included some form of explanation as to how the metrics used in fairness analyses had been calculated. Ten articles (3.2%) achieved this solely by providing the formula (as mathematical notation) for the metrics used in the article's fairness analysis, 77 articles (24.7%) provided a written description of the calculation process or interpretation of their selected metric's results and 178 articles (57.1%) provided both the formula and a written expiation for the calculation of the metrics used in each article's fairness analysis. Approximately one in ten articles (28, 9.0%) from our survey proposed a novel metric that they used to measure an aspect of fairness in a type of predictive algorithm.

The 312 articles in the analytic sample utilized 144 mathematically distinct metrics to assess fairness in algorithm-produced predictions. Table 6 shows that the median surveyed article's fairness analysis used 3 different metrics ($IQR = [1, 4]$) to measure fairness, with slightly less than half of all articles (143 articles, 45.8%) having used two or

Number of Articles	312
Article's Fairness Measure Selection Justification (%)	
Cite prior use	235 (75.3)
Evidence of validity	15 (4.8)
Plausible argument	32 (10.3)
Replication study	11 (3.5)
Article's Description of Fairness Measure (%)	
Formula and interpretation	178 (57.1)
Interpretation	77 (24.7)
Formula	10 (3.2)
None provided	47 (15.1)
Article Proposed New Fairness Measure (%)	
Yes	28 (9.0)
No	284 (91.0)

Table 5. Overview of justifications and descriptions of metrics used in fairness analyses applied to predictive algorithms 312 academic articles published between 2015 and 2025

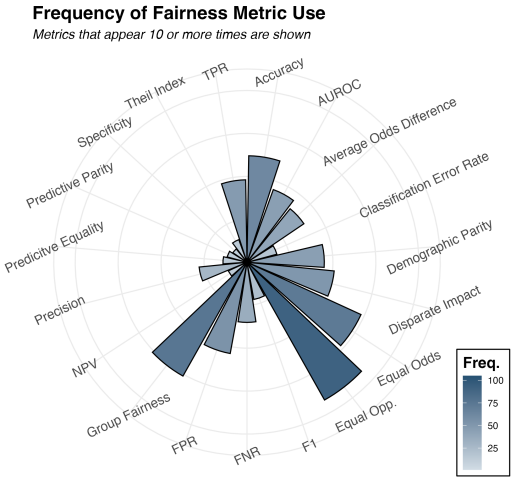

Fig. 3. Polar coordinate plot of the frequency of fairness metric use in the articles surveyed.

Number of Unique Fairness Measures	144
Confusion Matrix-Based Metrics	
n Unique (% of 144 Definitions)	60 (41.7)
Prevalence (% of 989 Uses)	787 (79.6)
Fairness Measures per Article	
Mean (SD)	3.17 (2.44)
Median [IQR]	3 [1,4]

Table 6. Overview of fairness metric use per article.

fewer metrics in their fairness analysis. Figure 3 is a polar coordinate plot representing the frequency of use of the 19 fairness metrics that had been used at least 10 times by articles in the survey. Sixty (41.7%) of the 144 unique metrics identified in this survey were based on binary classification confusion matrices. These confusion matrix-based metrics constituted four fifths of all metric uses across the 312 papers in our survey (787 of 8989 total metric uses, 79.6%). Additionally, approximately 10% of the unique metrics observed in the survey were specific to a certain domain such as difference in wait times in a health care setting or difference in blink sequence accuracy based in [vehicle] driver surveillance video [7, 17]

3.2.1 Authors's Interpretation: Fairness Measures and Metrics. The operationalizations of fairness on a metric level seen in the surveyed articles largely centres on achieving the egalitarian operationalization of fairness that we discussed in Section 3.1.1. This ethos can be seen in the high prevalence of metrics which define "fair" as a state where an algorithm's predictions meet an equality statement that compares prediction performance between groups or individuals in a data set. Importantly, the use of equality statements or acceptable thresholds for differences are not in-and-of themselves indicative of a metric being an egalitarian operationalization of fairness, in fact most operationalizations of fairness rely on these tools to some degree. It is rather the (often unstated) assumption that parity in performance between groups *is* fairness. In short, the vast majority of the metrics employed by the articles we surveyed are used in such a way that they functionally reduce the many complex and multi-faceted notions of fairness to `!bias == fair`.

4 Discussion

Having discussed the results of this literature survey study's sample, the fairness analyses observed, and the fairness metrics utilized, we now return to our initial research questions.

(1) *Which forms of fairness analysis are most prevalent in peer-reviewed literature that empirically assesses fairness in predictive algorithms using real-world data?*
Answer: More than 90% of the identified articles investigated binary classification problems. Over 85% of articles assessed fairness in a predictive algorithm by comparing it with another predictive algorithm while less than 3% articles assessed a predictive algorithm's fairness in comparison to human predictions.
   (a) *Which types of justifications were provided to support the selection of the fairness analysis methodologies implemented in the surveyed articles?*
   Answer: The most prevalent form of justification was to cite that the chosen fairness analysis method had been used previously in a different context without providing reasoning as to why that method would be valid in the article's context.

(2) *Which metrics are most frequently used to measure aspects of fairness within the fairness analyses of peer-reviewed articles that empirically assess predictive algorithms using real-world data?*
Answer: The most frequently used metrics in the articles surveyed were those derived from confusion matrices, such as equalized odds [19], as well as those derived solely from an algorithm's predictions, such as group fairness (AKA statistical parity difference) [10].
   (a) *Which types of justification were provided to support the context-specific validity of the fairness metrics used by these articles?*
   Answer: Three quarters of the surveyed articles cited the use of their chosen metrics generally, but neglected to provide evidence or reasoning that supported those metrics' validity in the context that the article had applied it to. Less than 5% of articles provided evidence that directly supported the context-specific validity of said article's selected metrics.

One of the central themes that emerged from this survey is the current relative lack of investigation into how to meaningfully measure predictive algorithm fairness. This is perhaps most clearly demonstrated by the 0 out of 329 articles from the survey that reported incorporating the perspectives of the people who the predictive algorithm was being applied to in their fairness analyses.

In line with the observations of previous surveys in this area, the vast majority of the articles in our survey investigated fairness in binary classification tasks using confusion matrix-based metrics [8, 24]. We also observed trends of reductively

defining fairness as a lack of bias, a general vagueness as to which aspect[s] of fairness were being assessed, and a lack of meaningful engagement with the real-world impacts of the predictive algorithms being assessed. Broadly, we would describe many of the articles in our survey as being written from a purely technical perspective, despite the sociotechnical nature of the problem being investigated.

While the themes we have highlighted so far in this discussion section represent high-level trends across the 329 articles surveyed, there many articles to which few or none of these generalizations apply. For example, Gerchick et al. [16] deeply engaged with the context of their fairness analysis as they investigated injustice and systematic discrimination in an algorithm used to predict the risk of child welfare neglect. Another example can be seen in the work of Imai et al. [21], which sought to develop a fairness measurement specifically for predictions made in pretrial Public Safety Assessments, a part of the US's criminal justice system.

Next steps for research in this area could include beginning to bridge the mathematical formulations of fairness utilized throughout the surveyed papers with notions of fairness grounded in the lived experience of those affected, or in any of the plethora of ethical and philosophical foundations that have been developed throughout the ages. Researchers could also begin establishing the context-specific validity of different approaches to fairness analysis in specific scenarios.

This survey has several strengths. To the best of our knowledge, this study is the first rigorous survey into how fairness analyses of predictive algorithms are used in practice. Accordingly, our survey builds upon the established knowledge base created by previous surveys which focused on aspects of fairness analyses such as identifying fairness metrics, creating taxonomies for bias and fairness-related terms, or summarizing current mitigation techniques [8, 24, 39]. This survey adds to this existing knowledge by providing context for the prevalence and frequency of use of those fairness analysis tools and metrics, as well as by providing insight into factors that could have led to the usage trends we observed. Further, as part of methodological rigour of our survey, we applied an adapted rapid review protocol and followed reporting best practices to enhance the accuracy and reproducibility of our results [14, 15]. This survey is also strengthened by the length of time from which articles were searched. Using a 10-year look back window allowed our survey to better contextualize the growth of research into predictive algorithm fairness over time. This survey also has several limitations. The practical decision to search the ACM library, SCOPUS, and CAS databases only means that papers which could have met our search criteria but were not indexed in those databases would have been missed. Given the varied and rapidly changing meanings of terms around fairness and predictive algorithms, (e.g., the seemingly ever-expanding definition of "AI" in the past 3 years [22]), our search criteria may not have sufficient captured all papers to which our meaning applied to if those papers did not use the terminology we developed. Knowing this would be a limitation, we consulted with a subject-matter expert librarian on our search terminology prior to conducting the survey. This survey is also limited in its conclusions due to the observational nature of surveys.

5 Conclusion

We began this search to better understand how researchers across disciplines are investigating fairness-related concerns in prediction algorithms that are being used on real-world data. Having now conducted the survey, we found a wide variety of methods and metrics that could be used in fairness analysis; however, we also found a general lack of engagement with how meaningfully the fairness assessments being used aligned with the real-world contexts that those analyses were being applied to. Particularly noticeable among the surveyed papers was the lack of inclusion of perspectives from the people affected by the algorithms being studied.

At its core, any assessment of fairness in predictive algorithms is an assessment of a sociotechnical phenomenon. Therefore, we believe it is important to underline that it is our responsibility as researchers to ensure that both the "socio" and the "technical" aspects of fairness in predictive algorithms are addressed in the analyses we conduct.

References

[1] Elizabeth S. Anderson. 1999. What Is the Point of Equality? *Ethics* 109, 2 (Jan. 1999), 287–337. jstor:10.1086/233897 doi:10.1086/233897

[2] Ashwathy Ashokan and Christian Haas. 2021. Fairness Metrics and Bias Mitigation Strategies for Rating Predictions. *Information Processing & Management* 58, 5 (Sept. 2021), N.PAG–N.PAG. doi:10.1016/j.ipm.2021.102646

[3] R.K.E. Bellamy, A. Mojsilovic, S. Nagar, K.N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K.R. Varshney, Y. Zhang, K. Dey, M. Hind, S.C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, and S. Mehta. 2019. AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias. *IBM Journal of Research and Development* 63, 4-5 (2019). doi:10.1147/JRD.2019.2942287

[4] Juliana Bidadanure and David Axelsen. 2025. Egalitarianism. In *The Stanford Encyclopedia of Philosophy* (spring 2025 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University.

[5] Justin B. Biddle, John P. Nelson, and Olajide E. Olugbade. 2024. How Can We Know If You Are Serious? Ethics Washing, Symbolic Ethics Offices, and the Responsible Design of AI Systems. *Canadian Journal of Philosophy* 54, 4 (May 2024), 329–345. doi:10.1017/can.2025.9

[6] Elettra Bietti. 2021. From Ethics Washing to Ethics Bashing: A Moral Philosophy View on Tech Ethics. *Journal of Social Computing* 2, 3 (Sept. 2021), 266–283. doi:10.23919/JSC.2021.0031

[7] M.M. Canellas, D.A. Pachamanova, G. Perakis, O. Skali Lami, and A. Tsiourvas. 2024. A Granular Approach to Optimal and Fair Patient Placement in Hospital Emergency Departments. *Production and Operations Management* (2024). doi:10.1177/10591478241240390

[8] Simon Caton and Christian Haas. 2024. Fairness in Machine Learning: A Survey. *Comput. Surveys* 56, 7 (July 2024), 1–38. doi:10.1145/3616865

[9] David M. Diez, Christopher D. Barr, and Mine Çetinkaya-Rundel. 2019. *OpenIntro Statistics* (fourth edition ed.). OpenIntro.

[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. doi:10.1145/2090236.2090255

[11] EBSCO Industries, LLC. 2025. Computers & Applied Sciences Complete. https://research.ebsco.com/c/udxg5x?db=iih.

[12] CSRC Content Editor. [n. d.]. Algorithm - Glossary | CSRC. https://csrc.nist.gov/glossary/term/algorithm.

[13] Elsevier B. V. 2025. Scopus. http://www.scopus.com/.

[14] Chantelle Garritty, Gerald Gartlehner, Barbara Nussbaumer-Streit, Valerie J. King, Candyce Hamel, Chris Kamel, Lisa Affengruber, and Adrienne Stevens. 2021. Cochrane Rapid Reviews Methods Group Offers Evidence-Informed Guidance to Conduct Rapid Reviews. *Journal of Clinical Epidemiology* 130 (Feb. 2021), 13–22. doi:10.1016/j.jclinepi.2020.10.007

[15] Chantelle Garritty, Candyce Hamel, Marialena Trivella, Gerald Gartlehner, Barbara Nussbaumer-Streit, Declan Devane, Chris Kamel, Ursula Griebler, and Valerie J. King. 2024. Updated Recommendations for the Cochrane Rapid Review Methods Guidance for Rapid Reviews of Effectiveness. *BMJ* 384 (Feb. 2024), e076335. doi:10.1136/bmj-2023-076335

[16] Marissa Gerchick, Tobi Jegede, Tarak Shah, Ana Gutierrez, Sophie Beiers, Noam Shemtov, Kath Xu, Anjana Samant, and Aaron Horowitz. 2023. The Devil Is in the Details: Interrogating Values Embedded in the Allegheny Family Screening Tool. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1292–1310. doi:10.1145/3593013.3594081

[17] Jakub Grzelak and Martim Brandao. 2021. The Dangers of Drowsiness Detection: Differential Performance, Downstream Impact, and Misuses. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (Aies '21)*. Association for Computing Machinery, New York, NY, USA, 525–531. doi:10.1145/3461702.3462593

[18] Michaela Hardt, Xiaoguang Chen, Xiaoyi Cheng, Michele Donini, Jason Gelman, Satish Gollaprolu, John He, Pedro Larroy, Xinyu Liu, Nick McCarthy, Ashish Rathi, Scott Rees, Ankit Siva, ErhYuan Tsai, Keerthan Vasist, Pinar Yilmaz, Muhammad Bilal Zafar, Sanjiv Das, Kevin Haas, Tyler Hill, and Krishnaram Kenthapadi. 2021. Amazon SageMaker Clarify: Machine Learning Bias Detection and Explainability in the Cloud. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (Kdd '21)*. Association for Computing Machinery, New York, NY, USA, 2974–2983. doi:10.1145/3447548.3467177

[19] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. doi:10.48550/ARXIV.1610.02413

[20] IBM Corporation. 2026. IBM Software Hub - Watson OpenScale. https://www.ibm.com/docs/en/software-hub/5.3.x?topic=services-watson-openscale.

[21] K. Imai, Z. Jiang, D.J. Greiner, R. Halen, and S. Shin. 2023. Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment. *Journal of the Royal Statistical Society. Series A: Statistics in Society* 186, 2 (2023), 167–189. doi:10.1093/jrsssa/qnad010

[22] Peter R. Lewis, Stephen Marsh, and Jeremy Pitt. 2021. AI vs "AI": Synthetic Minds or Speech Acts. *IEEE Technology and Society Magazine* 40, 2 (June 2021), 6–13. doi:10.1109/MTS.2021.3077052

[23] James Manyika and Sissie Hsiao. 2024. What Is Gemini and How It Works. https://gemini.google/overview/.

[24] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6 (July 2022), 1–35. doi:10.1145/3457607

[25] J. Misztal-Radecka and B. Indurkhya. 2023. A Bias Detection Tree Approach for Detecting Disparities in a Recommendation Model's Errors. *User Modeling and User-Adapted Interaction* 33, 1 (2023), 43–79. doi:10.1007/s11257-022-09334-x

[26] M. Moriah, F. Vermet, and A. Charpentier. 2024. Measuring and Mitigating Biases in Motor Insurance Pricing. *European Actuarial Journal* 14, 3 (2024), 833–869. doi:10.1007/s13385-024-00390-8

[27] Luke Munn. 2023. The Uselessness of AI Ethics. *AI and Ethics* 3, 3 (Aug. 2023), 869–877. doi:10.1007/s43681-022-00209-w

[28] OpenAI. [n. d.]. ChatGPT Overview. https://openai.com/chatgpt/overview/.

[29] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* 372 (March 2021), n71. doi:10.1136/bmj.n71

[30] Karl Pearson and Francis Galton. 1895. Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London* 58 (June 1895), 240–241.

[31] John Rawls. 2003. *A Theory of Justice* (rev. ed., 5.- 6. printing ed.). Belknap Press of Harvard Univ. Press, Cambridge, Mass.

[32] Michael J. Sandel. 2010. *Justice: What's the Right Thing to Do?* (1st pbk. ed ed.). Farrar, Straus and Giroux, New York.

[33] Laura Sikstrom, Marta M Maslej, Katrina Hui, Zoe Findlay, Daniel Z Buchman, and Sean L Hill. 2022. Conceptualising Fairness: Three Pillars for Medical Algorithms and Health Equity. *BMJ Health & Care Informatics* 29, 1 (Jan. 2022), e100459. doi:10.1136/bmjhci-2021-100459

[34] The Association for Computing Machinery. 2025. ACM Digital Library. https://dl.acm.org/.

[35] J. Ugirumurera, E.A. Bensen, J. Severino, and J. Sanyal. 2024. Addressing Bias in Bagging and Boosting Regression Models. *Scientific Reports* 14, 1 (2024). doi:10.1038/s41598-024-68907-5

[36] Gijs Van Maanen. 2022. AI Ethics, Ethics Washing, and the Need to Politicize Data Ethics. *Digital Society* 1, 2 (Sept. 2022), 9. doi:10.1007/s44206-022-00013-3

[37] Sriram Vasudevan and Krishnaram Kenthapadi. 2020. LiFT: A Scalable Framework for Measuring Fairness in ML Applications. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20)*. ACM. doi:10.1145/3340531.3412705

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.

[39] Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. In *Proceedings of the International Workshop on Software Fairness*. ACM, Gothenburg Sweden, 1–7. doi:10.1145/3194770.3194776

[40] Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. 2023. Fairlearn: Assessing and Improving Fairness of AI Systems. *Journal of Machine Learning Research* 24, 257 (2023), 1–8.

[41] Lindsay Weinberg. 2022. Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches. *Journal of Artificial Intelligence Research* 74 (May 2022), 75–109. doi:10.1613/jair.1.13196

A Search Strings

A.1 Generic Search String

(bias*) AND (fair* OR unfair* OR equit* OR justice OR prejudice OR discrimination) AND (algorithm OR “artificial intelligence” OR AI OR “machine learning” OR “machine-learning” OR ML OR “deep learning” OR “predictive model*” OR “predictive analy*”) AND (definition* OR outcome* OR metric* OR measure* OR framework OR test* OR assess* OR technique* OR tool* OR summar* OR eval* OR report*)

A.2 ACM Search

Search String:

[[Title: bias*] AND [[Title: fair*] OR [Title: unfair*] OR [Title: equit*] OR [Title: justice] OR [Title: prejudice] OR [Title: discrimination]] AND [[Title: algorithm] OR [Title: "artificial intelligence"] OR [Title: ai] OR [Title: "machine learning"] OR [Title: "machine-learning"] OR [Title: ml] OR [Title: "deep learning"] OR [Title: "predictive model*"] OR [Title: "predictive analy*"]] AND [[Title: definition*] OR [Title: outcome*] OR [Title: metric*] OR [Title: measure*] OR [Title: framework] OR [Title: test*] OR [Title: assess*] OR [Title: technique*] OR [Title: tool*] OR [Title: summar*] OR [Title: eval*] OR [Title: report*]]] OR [[Abstract: bias*] AND [[Abstract: fair*] OR [Abstract: unfair*] OR [Abstract: equit*] OR [Abstract: justice] OR [Abstract: prejudice] OR [Abstract: discrimination]] AND [[Abstract: algorithm] OR [Abstract: "artificial intelligence"] OR [Abstract: ai] OR [Abstract: "machine learning"] OR [Abstract: "machine-learning"] OR [Abstract: ml] OR [Abstract: "deep learning"] OR [Abstract: "predictive model*"] OR [Abstract: "predictive analy*"]] AND [[Abstract: definition*] OR [Abstract: outcome*] OR [Abstract: metric*] OR [Abstract: measure*] OR [Abstract: framework] OR [Abstract: test*] OR [Abstract: assess*] OR [Abstract: technique*] OR [Abstract: tool*] OR [Abstract: summar*] OR [Abstract: eval*] OR [Abstract: report*]]] OR [[Keywords: bias*] AND [[Keywords: fair*] OR [Keywords: unfair*] OR [Keywords: equit*] OR [Keywords: justice] OR [Keywords: prejudice] OR [Keywords: discrimination]] AND [[Keywords: algorithm] OR [Keywords: "artificial intelligence"] OR [Keywords: ai] OR [Keywords: "machine learning"] OR [Keywords: "machine-learning"] OR [Keywords: ml] OR [Keywords: "deep learning"] OR [Keywords: "predictive model*"] OR [Keywords: "predictive analy*"]] AND [[Keywords: definition*] OR [Keywords: outcome*] OR [Keywords: metric*] OR [Keywords: measure*] OR [Keywords: framework] OR [Keywords: test*] OR [Keywords: assess*] OR [Keywords: technique*] OR [Keywords: tool*] OR [Keywords: summar*] OR [Keywords: eval*] OR [Keywords: report*]]] AND [E-Publication Date: (01/01/2015 TO 31/12/2025)]

A.3 CAS Search

Limiters: Full Text; Peer Reviewed; Publication Date: 20150101-20251231. **Expanders:** Apply equivalent subjects. **Narrow by Language:** English.

Search String:

TI(("bias*") AND ("fair*" OR "unfair*" OR "equit*" OR "justice" OR "prejudice" OR "discrimination") AND ("algorithm" OR "artificial intelligence" OR "AI" OR "machine learning" OR "machine-learning" OR "ML" OR "deep learning" OR "predictive model*" OR "predictive analy*") AND ("definition*" OR

"outcome*" OR "metric*" OR "measure*" OR "framework" OR "test*" OR "assess*" OR "technique*" OR "tool*" OR "summar*" OR "eval*" OR "report*")) OR AB(("bias*" AND ("fair*" OR "unfair*" OR "equit*" OR "justice" OR "prejudice" OR "discrimination") AND ("algorithm" OR "artificial intelligence" OR "AI" OR "machine learning" OR "machine-learning" OR "ML" OR "deep learning" OR "predictive model*" OR "predictive analy*") AND ("definition*" OR "outcome*" OR "metric*" OR "measure*" OR "framework" OR "test*" OR "assess*" OR "technique*" OR "tool*" OR "summar*" OR "eval*" OR "report*")) OR KW(("bias*") AND ("fair*" OR "unfair*" OR "equit*" OR "justice" OR "prejudice" OR "discrimination") AND ("algorithm" OR "artificial intelligence" OR "AI" OR "machine learning" OR "machine-learning" OR "ML" OR "deep learning" OR "predictive model*" OR "predictive analy*") AND ("definition*" OR "outcome*" OR "metric*" OR "measure*" OR "framework" OR "test*" OR "assess*" OR "technique*" OR "tool*" OR "summar*" OR "eval*" OR "report*"))

A.4 SOCPUS Search

Limiters: Articles; Publication Year: > 2014 & < 2026. **Narrow by Language:** English.

Search String:

TITLE-ABS-KEY(("bias*") AND ("fair*" OR "unfair*" OR "equit*" OR "justice" OR "prejudice") AND ("algorithm" OR "artificial intelligence" OR "AI" OR "machine learning" OR "machine-learning" OR "deep learning" OR "predictive model*" OR "predictive analy*") AND ("metric*" OR "measure*" OR "framework" OR "test*" OR "assess*" OR "technique*" OR "tool*" OR "summar*" OR "eval*")) AND PUBYEAR > 2014 AND PUBYEAR < 2026 AND (LIMIT-TO(DOCTYPE,"ar")) AND (LIMIT-TO(LANGUAGE,"English"))

B Articles Included in Survey

Bibliography of Surveyed Articles

[42] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward Malthouse. 2021. User-Centered Evaluation of Popularity Bias in Recommender Systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (Umap '21)*. Association for Computing Machinery, New York, NY, USA, 119–129. doi:10.1145/3450613.3456821

[43] A. Abdulqawy, E. Sallam, and A. Elkholy. 2025. UnBias: Unveiling Bias Implications in Deep Learning Models for Healthcare Applications. *IEEE Journal of Biomedical and Health Informatics* 29, 2 (2025), 1464–1471. doi:10.1109/JBHI.2024.3484951

[44] Seyum Assefa Abebe, Claudio Lucchese, and Salvatore Orlando. 2022. EiFFFeL: Enforcing Fairness in Forests by Flipping Leaves. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing (Sac '22)*. Association for Computing Machinery, New York, NY, USA, 429–436. doi:10.1145/3477314.3507319

[45] Julian Abhari and Ashwin Ashok. 2023. Mitigating Racial Biases for Machine Learning Based Skin Cancer Detection. In *Proceedings of the Twenty-Fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc '23)*. Association for Computing Machinery, New York, NY, USA, 556–561. doi:10.1145/3565287.3617639

[46] M.D. Abràmoff, P.T. Lavin, J.R. Jakubowski, B.A. Blodi, M. Keeys, C. Joyce, and J.C. Folk. 2024. Mitigation of AI Adoption Bias through an Improved Autonomous AI System for Diabetic Retinal Disease. *npj Digital Medicine* 7, 1 (2024), 10 pages. doi:10.1038/s41746-024-01389-x

[47] E. Achterhold, M. Mühlböck, N. Steiber, and C. Kern. 2025. Fairness in Algorithmic Profiling: The AMAS Case. *Minds and Machines* 35, 1 (2025), 1–30. doi:10.1007/s11023-024-09706-9

[48] Halim Acosta, Nathan Henderson, Jonathan Rowe, Wookhee Min, James Minogue, and James Lester. 2021. What's Fair Is Fair: Detecting and Mitigating Encoded Bias in Multimodal Models of Museum Visitor Attention. In *Proceedings of the 2021 International Conference on Multimodal Interaction (Icmi '21)*. Association for Computing Machinery, New York, NY, USA, 258–267. doi:10.1145/3462244.3479943

[49] E. Ahn, Y. Gil, and E. Putnam-Hornstein. 2021. Predicting Youth at High Risk of Aging out of Foster Care Using Machine Learning Methods. *Child Abuse and Neglect* 117 (2021), 1–12. doi:10.1016/j.chiabu.2021.105059

[50] Nil-Jana Akpinar, Zachary Lipton, and Alexandra Chouldechova. 2024. The Impact of Differential Feature Under-Reporting on Algorithmic Fairness. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1355–1382. doi:10.1145/3630106.3658977

[51] A. Al-Zawqari, D. Peumans, and G. Vandersteen. 2024. Latent Space Bias Mitigation for Predicting At-Risk Students. *Computers and Education: Artificial Intelligence* 7 (2024), 1–15. doi:10.1016/j.caeai.2024.100300

[52] Mohammad Arif Ul Alam. 2021. AI-fairness towards Activity Recognition of Older Adults. In *MobiQuitous 2020 - 17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous '20)*. Association for Computing Machinery, New York, NY, USA, 108–117. doi:10.1145/3448891.3448943

[53] Jamal Alasadi, Ahmed Al Hilli, and Vivek K. Singh. 2019. Toward Fairness in Face Matching Algorithms. In *Proceedings of the 1st International Workshop on Fairness, Accountability, and Transparency in MultiMedia (Fat/Mm '19)*. Association for Computing Machinery, New York, NY, USA, 19–25. doi:10.1145/3347447.3356751

[54] V. Albiero, K. Zhang, M.C. King, and K.W. Bowyer. 2022. Gendered Differences in Face Recognition Accuracy Explained by Hairstyles, Makeup, and Facial Morphology. *IEEE Transactions on Information Forensics and Security* 17 (2022), 127–137. doi:10.1109/TIFS.2021.3135750

[55] C. Allen, V. Kumar, J. Elwell, S. Overman, B.S. Schoch, W. Aibinder, M. Parsons, J. Watling, J.K. Ko, B. Gobbato, T. Throckmorton, H. Routman, and C.P. Roche. 2024. Evaluating the Fairness and Accuracy of Machine Learning–Based Predictions of Clinical Outcomes after Anatomic and Reverse Total Shoulder Arthroplasty. *Journal of Shoulder and Elbow Surgery* 33, 4 (2024), 888–899. doi:10.1016/j.jse.2023.08.005

[56] Reem Alsaffar, Susan Gauch, Mohammed Alqahtani, and Omar Salman. 2024. Incorporating Fairness in Paper Recommendation. In *Proceedings of the 2021 ACM/IEEE Joint Conference on Digital Libraries (Jcdl '21)*. IEEE Press, Virtual Event, 318–319. doi:10.1109/JCDL52503.2021.00050

[57] C. Amaral, C. Kolsarici, and M. Nediak. 2024. Optimizing Pricing Delegation to External Sales Forces via Commissions: An Empirical Investigation. *Production and Operations Management* 33, 9 (2024), 1839–1854. doi:10.1177/10591478241259791

[58] E. Amigó, Y. Deldjoo, S. Mizzaro, and A. Bellogín. 2023. A Unifying and General Account of Fairness Measurement in Recommender Systems. *Information Processing and Management* 60, 1 (2023), 1–19. doi:10.1016/j.ipm.2022.103115

[59] H. Anahideh, N. Nezami, and A. Asudeh. 2025. Finding Representative Group Fairness Metrics Using Correlation Estimations. *Expert Systems with Applications* 262 (2025), 1–13. doi:10.1016/j.eswa.2024.125652

[60] E. Ascarza and A. Israeli. 2022. Eliminating Unintended Bias in Personalized Policies Using Bias-Eliminating Adapted Trees (BEAT). *Proceedings of the National Academy of Sciences of the United States of America* 119, 11 (2022), 1–9. doi:10.1073/pnas.2115293119

[61] Ashwathy Ashokan and Christian Haas. 2021. Fairness Metrics and Bias Mitigation Strategies for Rating Predictions. *Information Processing & Management* 58, 5 (Sept. 2021), N.PAG–N.PAG. doi:10.1016/j.ipm.2021.102646

[62] Pranjal Awasthi, Alex Beutel, Matthäus Kleindessner, Jamie Morgenstern, and Xuezhi Wang. 2021. Evaluating Fairness of Machine Learning Models under Uncertain and Incomplete Information. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 206–214. doi:10.1145/3442188.3445884

[63] G. Babaei and P. Giudici. 2024. How Fair Is Machine Learning in Credit Lending? *Quality and Reliability Engineering International* 40, 6 (2024), 3452–3464. doi:10.1002/qre.3579

[64] Youakim Badr and Rahul Sharma. 2022. Data Transparency and Fairness Analysis of the NYPD Stop-and-Frisk Program. *ACM Journal of Data and Information Quality* 14, 2, Article 7 (Feb. 2022), 1–14 pages. doi:10.1145/3460533

[65] Niels Bantilan. 2018. Themis-Ml: A Fairness-Aware Machine Learning Interface for End-To-End Discrimination Discovery and Mitigation. *Journal of Technology in Human Services* 36, 1 (Jan. 2018), 15–30. doi:10.1080/15228835.2017.1416512

[66] M. Benčević, M. Habijan, I. Galić, D. Babin, and A. Pižurica. 2024. Understanding Skin Color Bias in Deep Learning-Based Skin Lesion Segmentation. *Computer Methods and Programs in Biomedicine* 245 (2024), 1–10. doi:10.1016/j.cmpb.2024.108044

[67] J.A. Bergquist, B. Zenger, J. Brundage, R.S. MacLeod, T.J. Bunch, R. Shah, X. Ye, A. Lyons, M. Torre, R. Ranjan, T. Tasdizen, and B.A. Steinberg. 2024. Performance of Off-the-Shelf Machine Learning Architectures and Biases in Low Left Ventricular Ejection Fraction Detection. *Heart Rhythm O2* 5, 9 (2024), 644–654. doi:10.1016/j.hroo.2024.07.009

[68] R. Berk and A.A. Elzarka. 2020. Almost Politically Acceptable Criminal Justice Risk Assessment. *Criminology and Public Policy* 19, 4 (2020), 1231–1257. doi:10.1111/1745-9133.12500

[69] R.A. Berk, A.K. Kuchibhotla, and E. Tchetgen Tchetgen. 2024. Improving Fairness in Criminal Justice Algorithmic Risk Assessments Using Optimal Transport and Conformal Prediction Sets. *Sociological Methods and Research* 53, 4 (2024), 1629–1675. doi:10.1177/00491241231155883

[70] Karan Bhanot, Ioana Baldini, Dennis Wei, Jiaming Zeng, and Kristin Bennett. 2023. Stress-Testing Bias Mitigation Algorithms to Understand Fairness Vulnerabilities. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (Aies '23)*. Association for Computing Machinery, New York, NY, USA, 764–774. doi:10.1145/3600211.3604713

[71] K.A. Bhat and S.A. Sofi. 2024. Enhancing Machine Learning Efficacy and Fairness in Automated Decision Systems: An Adversarial Deep Generative Modeling with CoBS-TGAN Approach in Imbalanced and Biased Datasets. *International Journal of System Assurance Engineering and Management* (2024), 1–19. doi:10.1007/s13198-024-02430-8

[72] Abeba Birhane, Sepehr Dehdashtian, Vinay Prabhu, and Vishnu Boddeti. 2024. The Dark Side of Dataset Scaling: Evaluating Racial Classification in Multimodal Models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1229–1244. doi:10.1145/3630106.3658968

[73] Arpita Biswas, Marta Kolczynska, Saana Rantanen, and Polina Rozenshtein. 2020. The Role of In-Group Bias and Balanced Data: A Comparison of Human and Machine Recidivism Risk Predictions. In *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies (Compass '20)*. Association for Computing Machinery, New York, NY, USA, 97–104. doi:10.1145/3378393.3402507

[74] Sumon Biswas and Hridesh Rajan. 2020. Do the Machine Learning Models on a Crowd Sourced Platform Exhibit Bias? An Empirical Study on Model Fairness. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (Esec/Fse 2020)*. Association for Computing Machinery, New York, NY, USA, 642–653. doi:10.1145/3368089.3409704

[75] William Blanzeisky and Pádraig Cunningham. 2022. Using Pareto Simulated Annealing to Address Algorithmic Bias in Machine Learning. *Knowledge Engineering Review* 37 (Jan. 2022), 1–12. doi:10.1017/S0269888922000029

[76] C.H. Blow, L. Qian, C. Gibson, P. Obiomon, and X. Dong. 2024. Comprehensive Validation on Reweighting Samples for Bias Mitigation via AIF360. *Applied Sciences (Switzerland)* 14, 9 (2024), 1–16. doi:10.3390/app14093826

[77] L. Boeschoten, E.-J. van Kesteren, A. Bagheri, and D.L. Oberski. 2021. Achieving Fair Inference Using Error-Prone Outcomes. *International Journal of Interactive Multimedia and Artificial Intelligence* 6, 5 (2021), 9–15. doi:10.9781/ijimai.2021.02.007

[78] W.J. Bolton, R. Wilson, M. Gilchrist, P. Georgiou, A. Holmes, and T.M. Rawson. 2024. Personalising Intravenous to Oral Antibiotic Switch Decision Making through Fair Interpretable Machine Learning. *Nature Communications* 15, 1 (2024), 1–13. doi:10.1038/s41467-024-44740-2

[79] Rodrigo Borges and Kostas Stefanidis. 2022. F2VAE: A Framework for Mitigating User Unfairness in Recommendation Systems. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing (Sac '22)*. Association for Computing Machinery, New York, NY, USA, 1391–1398. doi:10.1145/3477314.3507152

[80] Dariusz Brzezinski, Julia Stachowiak, Jerzy Stefanowski, Izabela Szczech, Robert Susmaga, Sofya Aksenyuk, Uladzimir Ivashka, and Oleksandr Yasinskyi. 2024. Properties of Fairness Measures in the Context of Varying Class Imbalance and Protected Group Ratios. *ACM Transactions on Knowledge Discovery from Data* 18, 7 (Aug. 2024), 1–18. doi:10.1145/3654659

[81] T.S. Buda, J. Guerreiro, J. Omana Iglesias, C. Castillo, O. Smith, and A. Matic. 2022. Foundations for Fairness in Digital Health Apps. *Frontiers in Digital Health* 4 (2022), 1–15. doi:10.3389/fdgth.2022.943514

[82] Francois Buet-Golfouse and Islam Utyagulov. 2022. Towards Fair Unsupervised Learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1399–1409. doi:10.1145/3531146.3533197

[83] Muhammad Atif Butt, Adnan Qayyum, Hassan Ali, Ala Al-Fuqaha, and Junaid Qadir. 2023. Towards Secure Private and Trustworthy Human-Centric Embedded Machine Learning: An Emotion-Aware Facial Recognition Case Study. *Computers & Security* 125 (Feb. 2023), N.PAG–N.PAG. doi:10.1016/j.cose.2022.103058

[84] P. Cabanillas Silva, H. Sun, P. Rodriguez, M. Rezk, X. Zhang, J. Fliegenschmidt, N. Hulde, V. von Dossow, L. Meesseman, K. Depraetere, R. Szymanowsky, J. Stieg, and F.-M. Dahlweid. 2024. Evaluating Gender Bias in ML-based Clinical Risk Prediction Models: A Study on Multiple Use Cases at Different Hospitals. *Journal of Biomedical Informatics* 157 (2024), 1–8. doi:10.1016/j.jbi.2024.104692

[85] Kathleen Cachel and Elke Rundensteiner. 2022. FINS Auditing Framework: Group Fairness for Subset Selections. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (Aies '22)*. Association for Computing Machinery, New York, NY, USA, 144–155. doi:10.1145/3514094.3534160

[86] E.A. Campbell, S. Bose, and A.J. Masino. 2024. Conceptualizing Bias in EHR Data: A Case Study in Performance Disparities by Demographic Subgroups for a Pediatric Obesity Incidence Classifier. *PLOS Digital Health* 3, 10 (2024), 1–17. doi:10.1371/journal.pdig.0000642

[87] Ygor Canalli, Filipe Braida, Leandro Alvim, and Geraldo Zimbrão. 2024. Fair Transition Loss: From Label Noise Robustness to Bias Mitigation. *Knowledge-Based Systems* 294 (June 2024), N.PAG–N.PAG. doi:10.1016/j.knosys.2024.111711

[88] M.M. Canellas, D.A. Pachamanova, G. Perakis, O. Skali Lami, and A. Tsiourvas. 2024. A Granular Approach to Optimal and Fair Patient Placement in Hospital Emergency Departments. *Production and Operations Management* 34, 4 (2024), 575–589. doi:10.1177/10591478241240390

[89] S. Carbunaru, Y. Neshatvar, H. Do, K. Murray, R. Ranganath, and M. Nayan. 2024. Survival After Radical Cystectomy for Bladder Cancer: Development of a Fair Machine Learning Model. *JMIR Medical Informatics* 12 (2024), 1–12. doi:10.2196/63289

[90] S. Carbunaru, O.S. Nettey, P. Gogana, I.B. Helenowski, B. Jovanovic, M. Ruden, C.M.P. Hollowell, R. Sharifi, R.A. Kittles, E. Schaeffer, P. Gann, and A.B. Murphy. 2019. A Comparative Effectiveness Analysis of the PBCG vs. PCPT Risks Calculators in a Multi-Ethnic Cohort. *BMC Urology* 19, 1 (2019), 1–9. doi:10.1186/s12894-019-0553-6

[91] S. Caton, S. Malisetty, and C. Haas. 2022. Impact of Imputation Strategies on Fairness in Machine Learning. *Journal of Artificial Intelligence Research* 74 (2022), 1011–1035. doi:10.1613/jair.1.13197

[92] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. 2021. Bias in Machine Learning Software: Why? How? What to Do?. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (Esec/Fse 2021)*. Association for Computing Machinery, New York, NY, USA, 429–440. doi:10.1145/3468264.3468537

[93] Joymallya Chakraborty, Suvodeep Majumder, and Huy Tu. 2022. Fair-SSL: Building Fair ML Software with Less Data. In *Proceedings of the 2nd International Workshop on Equitable Data and Technology (FairWare '22)*. Association for Computing Machinery, New York, NY, USA, 1–8. doi:10.1145/3524491.3527305

[94] Joymallya Chakraborty, Suvodeep Majumder, Zhe Yu, and Tim Menzies. 2020. Fairway: A Way to Build Fair ML Software. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (Esec/Fse 2020)*. Association for Computing Machinery, New York, NY, USA, 654–665. doi:10.1145/3368089.3409697

[95] Eunice Chan, Zhining Liu, Ruizhong Qiu, Yuheng Zhang, Ross Maciejewski, and Hanghang Tong. 2024. Group Fairness via Group Consensus. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1788–1808. doi:10.1145/3630106.3659006

[96] C.-Y. Chang, J. Yuan, S. Ding, Q. Tan, K. Zhang, X. Jiang, X. Hu, and N. Zou. 2023. Towards Fair Patient-Trial Matching via Patient-Criterion Level Fairness Constraint. *AMIA Annual Symposium proceedings* 2023 (2023), 884–893.

[97] H. Chen, W. Lu, R. Song, and P. Ghosh. 2024. On Learning and Testing of Counterfactual Fairness through Data Preprocessing. *J. Amer. Statist. Assoc.* 119, 546 (2024), 1286–1296. doi:10.1080/01621459.2023.2186885

[98] H. Chen, T. Zhu, B. Liu, W. Zhou, and P.S. Yu. 2024. Fine-Tuning a Biased Model for Improving Fairness. *IEEE Transactions on Big Data* 11, 3 (2024), 1397–1410. doi:10.1109/TBDATA.2024.3460537

[99] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. 2023. A Comprehensive Empirical Study of Bias Mitigation Methods for Machine Learning Classifiers. *ACM Transactions on Software Engineering and Methodology* 32, 4, Article 106 (May 2023), 30 pages. doi:10.1145/3583561

[100] Hao-Fei Cheng, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (Chi '22)*. Association for Computing Machinery, New York, NY, USA, Article 162, 22 pages. doi:10.1145/3491102.3501831

[101] Lingwei Cheng, Cameron Drayton, Alexandra Chouldechova, and Rhema Vaithianathan. 2025. Algorithm-Assisted Decision Making and Racial Disparities in Housing: A Study of the Allegheny Housing Assessment Tool. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (Aies '24)*. AAAI Press, San Jose, California, USA, 281–292.

[102] C.-H. Chiu, Y.-J. Chen, Y. Wu, Y. Shi, and T.-Y. Ho. 2024. Achieve Fairness without Demographics for Dermatological Disease Diagnosis. *Medical Image Analysis* 95 (2024), 1–13. doi:10.1016/j.media.2024.103188

[103] Jaeyoon Choi, Shamya Karumbaiah, and Jeffrey Matayoshi. 2025. Bias or Insufficient Sample Size? Improving Reliable Estimation of Algorithmic Bias for Minority Groups. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference (Lak '25)*. Association for Computing Machinery, New York, NY, USA, 547–557. doi:10.1145/3706468.3706540

[104] Jade Mai Cock, Muhammad Bilal, Richard Davis, Mirko Marras, and Tanja Kaser. 2023. Protected Attributes Tell Us Who, Behavior Tells Us How: A Comparison of Demographic and Behavioral Oversampling for Fair Student Success Modeling. In *LAK23: 13th International Learning Analytics and Knowledge Conference (LAK2023)*. Association for Computing Machinery, New York, NY, USA, 488–498. doi:10.1145/3576050.3576149

[105] J. Coe and M. Atay. 2021. Evaluating Impact of Race in Facial Recognition across Machine Learning and Deep Learning Algorithms. *Computers* 10, 9 (2021), 1–24. doi:10.3390/computers10090113

[106] Ivona Colakovic and Sašo Karakatič. 2023. FairBoost: Boosting Supervised Learning for Learning on Multiple Sensitive Features. *Knowledge-Based Systems* 280 (Nov. 2023), N.PAG–N.PAG. doi:10.1016/j.knosys.2023.110999

[107] S.P. Colegate, A. Palipana, E. Gecili, R.D. Szczesniak, and C. Brokamp. 2024. Evaluating Precision Medicine Tools in Cystic Fibrosis for Racial and Ethnic Fairness. *Journal of Clinical and Translational Science* 8, 1 (2024), 1–7. doi:10.1017/cts.2024.532

[108] J.P. Consuegra-Ayala, Y. Gutiérrez, Y. Almeida-Cruz, and M. Palomar. 2025. Bias Mitigation for Fair Automation of Classification Tasks. *Expert Systems* 42, 2 (2025), 1–17. doi:10.1111/exsy.13734

[109] A. Corbin and O. Marques. 2023. Assessing Bias in Skin Lesion Classifiers with Contemporary Deep Learning and Post-Hoc Explainability Techniques. *IEEE Access* 11 (2023), 78339–78352. doi:10.1109/ACCESS.2023.3289320

[110] Giandomenico Cornacchia, Vito Walter Anelli, Giovanni Maria Biancofiore, Fedelucio Narducci, Claudio Pomo, Azzurra Ragone, and Eugenio Di Sciascio. 2023. Auditing Fairness under Unawareness through Counterfactual Reasoning. *Information Processing & Management* 60, 2 (March 2023), N.PAG–N.PAG. doi:10.1016/j.ipm.2022.103224

[111] R. Correa, K. Pahwa, B. Patel, C.M. Vachon, J.W. Gichoya, and I. Banerjee. 2024. Efficient Adversarial Debiasing with Concept Activation Vector — Medical Image Case-Studies. *Journal of Biomedical Informatics* 149 (2024), 1–8. doi:10.1016/j.jbi.2023.104548

[112] S. D. Almeida, T. Norajitra, C.T. Lüth, T. Wald, V. Weru, M. Nolden, P.F. Jäger, O. von Stackelberg, C.P. Heußel, O. Weinheimer, J. Biederer, H.-U. Kauczor, and K. Maier-Hein. 2024. How Do Deep-Learning Models Generalize across Populations? Cross-ethnicity Generalization of COPD Detection. *Insights into Imaging* 15, 1 (2024), 1–12. doi:10.1186/s13244-024-01781-x

[113] S. Dakshit, S. Dakshit, N. Khargonkar, and B. Prabhakaran. 2023. Bias Analysis in Healthcare Time Series (BAHT) Decision Support Systems from Meta Data. *Journal of Healthcare Informatics Research* 7, 2 (2023), 225–253. doi:10.1007/s41666-023-00133-6

[114] Giordano d'Aloisio, Andrea D'Angelo, Antinisca Di Marco, and Giovanni Stilo. 2023. Debiaser for Multiple Variables to Enhance Fairness in Classification Tasks. *Information Processing & Management* 60, 2 (March 2023), N.PAG–N.PAG. doi:10.1016/j.ipm.2022.103226

[115] V.N. Dang, A. Cascarano, R.H. Mulder, C. Cecil, M.A. Zuluaga, J. Hernández-González, and K. Lekadir. 2024. Fairness and Bias Correction in Machine Learning for Depression Prediction across Four Study Populations. *Scientific Reports* 14, 1 (2024), 1–12. doi:10.1038/s41598-024-58427-7

[116] Vishnu Asutosh Dasu, Ashish Kumar, Saeid Tizpaz-Niari, and Gang Tan. 2024. NeuFair: Neural Network Fairness Repair with Dropout. In *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis (Issta 2024)*. Association for Computing Machinery, New York, NY, USA, 1541–1553. doi:10.1145/3650212.3680380

[117] A. Davoudi, S. Chae, L. Evans, S. Sridharan, J. Song, K.H. Bowles, M.V. McDonald, and M. Topaz. 2024. Fairness Gaps in Machine Learning Models for Hospitalization and Emergency Department Visit Risk Prediction in Home Healthcare Patients with Heart Failure. *International Journal of Medical Informatics* 191 (2024), 1–7. doi:10.1016/j.ijmedinf.2024.105534

[118] A. Davoudi, R. Sajdeya, R. Ison, J. Hagen, P. Rashidi, C.C. Price, and P.J. Tighe. 2023. Fairness in the Prediction of Acute Postoperative Pain Using Machine Learning Models. *Frontiers in Digital Health* 4 (2023), 1–12. doi:10.3389/fdgth.2022.970281

[119] Luiz Fernando F. P. de Lima, Danielle Rousy D. Ricarte, and Clauirton A. Siebra. 2024. A Benchmark Proposal for Non-Generative Fair Adversarial Learning Strategies Using a Fairness-Utility Trade-off Metric. *Computational Intelligence* 40, 5 (Oct. 2024), 1–20. doi:10.1111/coin.70003

[120] M. DeMichele, P. Baumgartner, M. Wenger, K. Barrick, and M. Comfort. 2020. Public Safety Assessment: Predictive Utility and Differential Prediction by Race in Kentucky. *Criminology and Public Policy* 19, 2 (2020), 409–431. doi:10.1111/1745-9133.12481

[121] Cyrus DiCiccio, Sriram Vasudevan, Kinjal Basu, Krishnaram Kenthapadi, and Deepak Agarwal. 2020. Evaluating Fairness Using Permutation Tests. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (Kdd '20)*. Association for Computing Machinery, New York, NY, USA, 1467–1477. doi:10.1145/3394486.3403199

[122] S. Dietrich, D. Malerba, and F. Gassmann. 2024. Predicting Social Assistance Beneficiaries: On the Social Welfare Damage of Data Biases. *Data and Policy* 6 (2024), 28 pages. doi:10.1017/dap.2023.38

[123] V. Digumarthi, T. Amin, S. Kanu, J. Mathew, B. Edwards, L.A. Peterson, M.E. Lundy, and K.E. Hegarty. 2024. Preoperative Prediction Model for Risk of Readmission after Total Joint Replacement Surgery: A Random Forest Approach Leveraging NLP and Unfairness Mitigation for Improved Patient Care and Cost-Effectiveness. *Journal of Orthopaedic Surgery and Research* 19, 1 (2024), 1–17. doi:10.1186/s13018-024-04774-0

[124] Yasmine Djebrouni, Nawel Benarba, Ousmane Touat, Pasquale De Rosa, Sara Bouchenak, Angela Bonifati, Pascal Felber, Vania Marangozova, and Valerio Schiavoni. 2024. Bias Mitigation in Federated Learning for Edge Computing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4, Article 157 (Jan. 2024), 35 pages. doi:10.1145/3631455

[125] H. Do, S. Nandi, P. Putzel, P. Smyth, and J. Zhong. 2023. A Joint Fairness Model with Applications to Risk Predictions for Underrepresented Populations. *Biometrics* 79, 2 (2023), 826–840. doi:10.1111/biom.13632

[126] I. Dominguez-Catena, D. Paternain, A. Jurio, and M. Galar. 2024. Less Can Be More: Representational vs. Stereotypical Gender Bias in Facial Expression Recognition. *Progress in Artificial Intelligence* 14, 11 (2024), 1–21. doi:10.1007/s13748-024-00345-w

[127] J. Dressel and H. Farid. 2018. The Accuracy, Fairness, and Limits of Predicting Recidivism. *Science Advances* 4, 1 (2018), 1–5. doi:10.1126/sciadv.aao5580

[128] Chengwen Du and Tao Chen. 2024. Contexts Matter: An Empirical Study on Contextual Influence in Fairness Testing for Deep Learning Systems. In *Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (Esem '24)*. Association for Computing Machinery, New York, NY, USA, 107–118. doi:10.1145/3674805.3686673

[129] David Quashigah Dzakpasu, Jixue Liu, Jiuyong Li, and Lin Liu. 2024. Integrating Fair Representation Learning with Fairness Regularization for Intersectional Group Fairness. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (Cikm '24)*. Association for Computing Machinery, New York, NY, USA, 560–569. doi:10.1145/3627673.3679802

[130] A. Erfani and V. Frias-Martinez. 2023. A Fairness Assessment of Mobility-Based COVID-19 Case Prediction Models. *PLoS ONE* 18, 10 (2023), 1–16. doi:10.1371/journal.pone.0292090

[131] A. Esmaelizadeh, S. Cotterill, L. Hebert, L. Golab, and K. Taghva. 2025. Infomod: Information-Theoretic Machine Learning Model Diagnostics. *Distributed and Parallel Databases* 43, 1 (2025), 1–34. doi:10.1007/s10619-024-07450-8

[132] Hossein Estiri, Zachary H Strasser, Sina Rashidian, Jeffrey G Klann, Kavishwar B Wagholikar, Thomas H McCoy, and Shawn N Murphy. 2022. An Objective Framework for Evaluating Unrecognized Bias in Medical AI Models Predicting COVID-19 Outcomes. *Journal of the American Medical Informatics Association* 29, 8 (Aug. 2022), 1334–1341. doi:10.1093/jamia/ocac070

[133] M. Fang, W. Yang, A. Kuijper, V. Štruc, and N. Damer. 2024. Fairness in Face Presentation Attack Detection. *Pattern Recognition* 147 (2024), 16. doi:10.1016/j.patcog.2023.110002

[134] Michael Mayowa Farayola, Malika Bendechache, Takfarinas Saber, Regina Connolly, and Irina Tal. 2024. Enhancing Algorithmic Fairness: Integrative Approaches and Multi-Objective Optimization Application in Recidivism Models. In *Proceedings of the 19th International Conference on Availability, Reliability and Security (Ares '24)*. Association for Computing Machinery, New York, NY, USA, Article 114, 10 pages. doi:10.1145/3664476.3669978

[135] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Kdd '15)*. Association for Computing Machinery, New York, NY, USA, 259–268. doi:10.1145/2783258.2783311

[136] Martínez-Plumed Fernando, Ferri Cèsar, Nieves David, and Hernández-Orallo José. 2021. Missing the Missing Values: The Ugly Duckling of Fairness in Machine Learning. *International Journal of Intelligent Systems* 36, 7 (July 2021), 3217–3258. doi:10.1002/int.22415

[137] Julien Ferry, Ulrich Aïvodji, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. 2023. Improving Fairness Generalization through a Sample-Robust Optimization Method. *Machine Learning* 112, 6 (June 2023), 2131–2192. doi:10.1007/s10994-022-06191-y

[138] R.R. Fletcher, A. Nakeshimana, and O. Olubeko. 2021. Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health. *Frontiers in Artificial Intelligence* 3 (2021), 1–17. doi:10.3389/frai.2020.561802

[139] Sainyam Galhotra, Karthikeyan Shanmugam, Prasanna Sattigeri, and Kush R. Varshney. 2022. Causal Feature Selection for Algorithmic Fairness. In *Proceedings of the 2022 International Conference on Management of Data (Sigmod '22)*. Association for Computing Machinery, New York, NY, USA, 276–285. doi:10.1145/3514221.3517909

[140] D. Gándara, H. Anahideh, M.P. Ison, and L. Picchiarini. 2024. Inside the Black Box: Detecting and Mitigating Algorithmic Bias Across Racialized Groups in College Student-Success Prediction. *AERA Open* 10 (2024), 1–15. doi:10.1177/23328584241258741

[141] M.P. Gangan, A. Kadan, and V.L. Lajish. 2024. Toward Exploring Fairness in Visual Transformer Based Natural and GAN Image Detection Systems. *IEEE Transactions on Computational Social Systems* 12, 3 (2024), 1068–1079. doi:10.1109/TCSS.2024.3509340

[142] T. Ganta, A. Kia, P. Parchure, M.-H. Wang, M. Besculides, M. Mazumdar, and C.B. Smith. 2024. Fairness in Predicting Cancer Mortality Across Racial Subgroups. *JAMA Network Open* 7, 7 (2024), e2421290. doi:10.1001/jamanetworkopen.2024.21290

[143] J. Gao, P. Mar, Z.-Z. Tang, and G. Chen. 2024. Fair Prediction of 2-Year Stroke Risk in Patients with Atrial Fibrillation. *Journal of the American Medical Informatics Association* 31, 12 (2024), 2820–2828. doi:10.1093/jamia/ocae170

[144] D. García-Soriano and F. Bonchi. 2020. Fair-by-Design Matching. *Data Mining and Knowledge Discovery* 34, 5 (2020), 1291–1335. doi:10.1007/s10618-020-00675-y

[145] A. Gebran, S.S. Thakur, L.R. Maurer, H. Bandi, R. Sinyard, A. Dorken-Gallastegi, M. Bokenkamp, M. El Moheb, L. Naar, A. Vapsi, D. Daye, G.C. Velmahos, D. Bertsimas, and H.M.A. Kaafarani. 2023. Development of a Machine Learning-Based Prescriptive Tool to Address Racial Disparities in Access to Care after Penetrating Trauma. *JAMA Surgery* 158, 10 (2023), 1088–1095. doi:10.1001/jamasurg.2023.2293

[146] Marissa Gerchick, Tobi Jegede, Tarak Shah, Ana Gutierrez, Sophie Beiers, Noam Shemtov, Kath Xu, Anjana Samant, and Aaron Horowitz. 2023. The Devil Is in the Details: Interrogating Values Embedded in the Allegheny Family Screening Tool. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1292–1310. doi:10.1145/3593013.3594081

[147] B. Ghai and K. Mueller. 2023. D-BIAS: A Causality-Based Human-in-the-Loop System for Tackling Algorithmic Bias. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2023), 473–482. doi:10.1109/TVCG.2022.3209484

[148] Azin Ghazimatin, Matthaus Kleindessner, Chris Russell, Ziawasch Abedjan, and Jacek Golebiowski. 2022. Measuring Fairness of Rankings under Noisy Sensitive Information. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2263–2279. doi:10.1145/3531146.3534641

[149] Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S. Meel. 2023. "How Biased Are Your Features?": Computing Fairness Influence Functions with Global Sensitivity Analysis. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 138–148. doi:10.1145/3593013.3593983

[150] A. Giloni, E. Grolman, T. Hagemann, R. Fromm, S. Fischer, Y. Elovici, and A. Shabtai. 2024. BENN: Bias Estimation Using a Deep Neural Network. *IEEE Transactions on Neural Networks and Learning Systems* 35, 1 (2024), 117–131. doi:10.1109/TNNLS.2022.3172365

[151] Sofie Goethals, David Martens, and Toon Calders. 2024. PreCoF: Counterfactual Explanations for Fairness. *Machine Learning* 113, 5 (May 2024), 3111–3142. doi:10.1007/s10994-023-06319-8

[152] Usman Gohar, Sumon Biswas, and Hridesh Rajan. 2023. Towards Understanding Fairness and Its Composition in Ensemble Machine Learning. In *Proceedings of the 45th International Conference on Software Engineering (Icse '23)*. IEEE Press, Melbourne, Victoria, Australia, 1533–1545. doi:10.1109/ICSE48619.2023.00133

[153] V. Grari, B. Ruf, S. Lamprier, and M. Detyniecki. 2020. Achieving Fairness with Decision Trees: An Adversarial Approach. *Data Science and Engineering* 5, 2 (2020), 99–110. doi:10.1007/s41019-020-00124-2

[154] Kylie Griep, Leah Stone, Stephen Edwards, Malik Hill, Jialin Cui, Chengyuan Liu, Qinjin Jia, and Yang Song. 2024. Ensuring Ethical, Transparent, and Auditable Use of Education Data and Algorithms on AutoML. In *Proceedings of the 2023 6th International Conference on Machine Learning and Natural Language Processing (Mlnlp '23)*. Association for Computing Machinery, New York, NY, USA, 66–72. doi:10.1145/3639479.3639492

[155] M. Groh, O. Badri, R. Daneshjou, A. Koochek, C. Harris, L.R. Soenksen, P.M. Doraiswamy, and R. Picard. 2024. Deep Learning-Aided Decision Support for Diagnosis of Skin Disease across Skin Tones. *Nature Medicine* 30, 2 (2024), 573–583. doi:10.1038/s41591-023-02728-3

[156] Jakub Grzelak and Martim Brandao. 2021. The Dangers of Drowsiness Detection: Differential Performance, Downstream Impact, and Misuses. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (Aies '21)*. Association for Computing Machinery, New York, NY, USA, 525–531. doi:10.1145/3461702.3462593

[157] T. Gu, W. Pan, J. Yu, G. Ji, X. Meng, Y. Wang, and M. Li. 2025. Mitigating Bias in AI Mortality Predictions for Minority Populations: A Transfer Learning Approach. *BMC Medical Informatics and Decision Making* 25, 1 (2025), 1–11. doi:10.1186/s12911-025-02862-7

[158] Vitória Guardieiro, Marcos M. Raimundo, and Jorge Poco. 2023. Enforcing Fairness Using Ensemble of Diverse Pareto-optimal Models. *Data Mining & Knowledge Discovery* 37, 5 (Sept. 2023), 1930–1958. doi:10.1007/s10618-023-00922-y

[159] D. Guo, C. Wang, B. Wang, and H. Zha. 2024. Learning Fair Representations via Distance Correlation Minimization. *IEEE Transactions on Neural Networks and Learning Systems* 35, 2 (2024), 2139–2152. doi:10.1109/TNNLS.2022.3187165

[160] A. Gupta, D. Carpenter, W. Min, J. Rowe, R. Azevedo, and J. Lester. 2024. Detecting and Mitigating Encoded Bias in Deep Learning-Based Stealth Assessment Models for Reflection-Enriched Game-Based Learning Environments. *International Journal of Artificial Intelligence in Education* 34, 3 (2024), 1138–1165. doi:10.1007/s40593-023-00379-6

[161] Aparajita Haldar, Teddy Cunningham, and Hakan Ferhatosmanoglu. 2022. RAGUEL: Recourse-aware Group Unfairness Elimination. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (Cikm '22)*. Association for Computing Machinery, New York, NY, USA, 666–675. doi:10.1145/3511808.3557424

[162] Mengde Han, Tianqing Zhu, and Wanlei Zhou. 2024. Fair Federated Learning with Opposite GAN. *Knowledge-Based Systems* 287 (March 2024), N.PAG–N.PAG. doi:10.1016/j.knosys.2024.111420

[163] C.A. Hane and M. Wasserman. 2023. Designing Equitable Health Care Outreach Programs From Machine Learning Patient Risk Scores. *Medical Care Research and Review* 80, 2 (2023), 216–227. doi:10.1177/10775587221098831

[164] Michaela Hardt, Xiaoguang Chen, Xiaoyi Cheng, Michele Donini, Jason Gelman, Satish Gollaprolu, John He, Pedro Larroy, Xinyu Liu, Nick McCarthy, Ashish Rathi, Scott Rees, Ankit Siva, ErhYuan Tsai, Keerthan Vasist, Pinar Yilmaz, Muhammad Bilal Zafar, Sanjiv Das, Kevin Haas, Tyler Hill, and Krishnaram Kenthapadi. 2021. Amazon SageMaker Clarify: Machine Learning Bias Detection and Explainability in the Cloud. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (Kdd '21)*. Association for Computing Machinery, New York, NY, USA, 2974–2983. doi:10.1145/3447548.3467177

[165] Emma Harvey, Michelle Seng Ah Lee, and Jatinder Singh. 2024. Improving Group Fairness Assessments with Proxies. *ACM Journal on Responsible Computing* 1, 4, Article 23 (Nov. 2024), 21 pages. doi:10.1145/3677175

[166] Erhu He, Yiqun Xie, Xiaowei Jia, Weiye Chen, Han Bao, Xun Zhou, Zhe Jiang, Rahul Ghosh, and Praveen Ravirathinam. 2022. Sailing in the Location-Based Fairness-Bias Sphere. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems (Sigspatial '22)*. Association for Computing Machinery, New York, NY, USA, Article 41, 10 pages. doi:10.1145/3557915.3560976

[167] Fang He, Xiaoxia Wu, Wenyu Zhang, and Xiaoling Huang. 2023. A Novel Bias-Alleviated Hybrid Ensemble Model Based on over-Sampling and Post-Processing for Fair Classification. *Connection Science* 35, 1 (Dec. 2023), 1–20. doi:10.1080/09540091.2023.2184310

[168] Yuzi He, Keith Burghardt, and Kristina Lerman. 2020. A Geometric Solution to Fair Representations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (Aies '20)*. Association for Computing Machinery, New York, NY, USA, 279–285. doi:10.1145/3375627.3375864

[169] Z. He, Z. Wang, X. Dong, P. Sun, J. Ren, and K. Ren. 2025. Towards Fair Federated Learning via Unbiased Feature Aggregation. *IEEE Transactions on Dependable and Secure Computing* 22 (2025), 3795–3807. Issue 4. doi:10.1109/TDSC.2025.3541306

[170] L. Herrmann and J. Weigert. 2024. AI-based Prediction of Academic Success: Support for Many, Disadvantage for Some? *Computers and Education: Artificial Intelligence* 7 (2024), 1–7. doi:10.1016/j.caeai.2024.100303

[171] C. Hong, M. Liu, D.M. Wojdyla, J. Hickey, M. Pencina, and R. Henao. 2024. Trans-Balance: Reducing Demographic Disparity for Prediction Models in the Presence of Class Imbalance. *Journal of Biomedical Informatics* 149 (2024), 1–10. doi:10.1016/j.jbi.2023.104532

[172] Junyuan Hong, Zhuangdi Zhu, Shuyang Yu, Zhangyang Wang, Hiroko H. Dodge, and Jiayu Zhou. 2021. Federated Adversarial Debiasing for Fair and Transferable Representations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (Kdd '21)*. Association for Computing Machinery, New York, NY, USA, 617–627. doi:10.1145/3447548.3467281

[173] Max Hort and Federica Sarro. 2022. Privileged and Unprivileged Groups: An Empirical Study on the Impact of the Age Attribute on Fairness. In *Proceedings of the 2nd International Workshop on Equitable Data and Technology (FairWare '22)*. Association for Computing Machinery, New York, NY, USA, 17–24. doi:10.1145/3524491.3527308

[174] M. Hort, J.M. Zhang, F. Sarro, and M. Harman. 2024. Search-Based Automatic Repair for Fairness and Accuracy in Decision-making Software. *Empirical Software Engineering* 29, 1 (2024), 1–33. doi:10.1007/s10664-023-10419-3

[175] Max Hort, Jie M. Zhang, Federica Sarro, and Mark Harman. 2021. Fairea: A Model Behaviour Mutation Approach to Benchmarking Bias Mitigation Methods. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (Esec/Fse 2021)*. Association for Computing Machinery, New York, NY, USA, 994–1006. doi:10.1145/3468264.3468565

[176] L. Hu, D. Li, H. Liu, X. Chen, Y. Gao, S. Huang, X. Peng, X. Zhang, X. Bai, H. Yang, L. Kong, J. Tang, P. Lu, C. Xiong, and H. Liang. 2024. Enhancing Fairness in AI-enabled Medical Systems with the Attribute Neutral Framework. *Nature Communications* 15, 1 (2024), 1–16. doi:10.1038/s41467-024-52930-1

[177] Y. Hu, T. Liao, J. Chen, J. Bian, Z. Zheng, and C. Chen. 2024. Migrate Demographic Group for Fair Graph Neural Networks. *Neural Networks* 175 (2024), 1–12. doi:10.1016/j.neunet.2024.106264

[178] J.A. Idowu, A.S. Koshiyama, and P. Treleaven. 2024. Investigating Algorithmic Bias in Student Progress Monitoring. *Computers and Education: Artificial Intelligence* 7 (2024), 1–13. doi:10.1016/j.caeai.2024.100267

[179] K. Imai, Z. Jiang, D.J. Greiner, R. Halen, and S. Shin. 2023. Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment. *Journal of the Royal Statistical Society. Series A: Statistics in Society* 186, 2 (2023), 167–189. doi:10.1093/jrsssa/qnad010

[180] Vasileios Iosifidis, Arjun Roy, and Eirini Ntoutsi. 2022. Parity-Based Cumulative Fairness-Aware Boosting. *Knowledge & Information Systems* 64, 10 (Oct. 2022), 2737–2770. doi:10.1007/s10115-022-01723-3

[181] Maliha Tashfia Islam, Anna Fariha, Alexandra Meliou, and Babak Salimi. 2022. Through the Data Management Lens: Experimental Analysis and Evaluation of Fair Classification. In *Proceedings of the 2022 International Conference on Management of Data (Sigmod '22)*. Association for Computing Machinery, New York, NY, USA, 232–246. doi:10.1145/3514221.3517841

[182] L. Iurada, S. Bucci, T.M. Hospedales, and T. Tommasi. 2024. Fairness Meets Cross-Domain Learning: A Benchmark of Models and Metrics. *IEEE Access* 12 (2024), 47854–47867.

[183] B. Jain, M. Huber, R. Elmasri, and L. Fegaras. 2020. Using Bias Parity Score to Find Feature-Rich Models with Least Relative Bias. *Technologies* 8, 4 (2020), 1–21. doi:10.3390/technologies8040068

[184] Bhanu Jain, Manfred Huber, Ramez A. Elmasri, and Leonidas Fegaras. 2020. Reducing Race-Based Bias and Increasing Recidivism Prediction Accuracy by Using Past Criminal History Details. In *Proceedings of the 13th ACM International Conference on Pervasive Technologies Related to Assistive Environments (Petra '20)*. Association for Computing Machinery, New York, NY, USA, Article 55, 8 pages. doi:10.1145/3389189.3397990

[185] Bhanu Jain, Manfred Huber, Leonidas Fegaras, and Ramez A. Elmasri. 2019. Singular Race Models: Addressing Bias and Accuracy in Predicting Prisoner Recidivism. In *Proceedings of the 12th ACM International Conference on Pervasive Technologies Related to Assistive Environments (Petra '19)*. Association for Computing Machinery, New York, NY, USA, 599–607. doi:10.1145/3316782.3322787

[186] T. Janevic, L.E. Tomalin, K.B. Glazer, N. Boychuk, A. Kern-Goldberger, M. Burdick, F. Howell, M. Suarez-Farinas, N. Egorova, J. Zeitlin, P. Hebert, and E.A. Howell. 2024. Development of a Prediction Model of Postpartum Hospital Use Using an Equity-Focused Approach. *American Journal of Obstetrics and Gynecology* 230, 6 (2024), 671.e1–671.e10. doi:10.1016/j.ajog.2023.10.033

[187] J.E. Johndrow and K. Lum. 2019. An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction. *Annals of Applied Statistics* 13, 1 (2019), 189–220. doi:10.1214/18-AOAS1201

[188] M.S. Johnson, X. Liu, and D.F. McCaffrey. 2022. Psychometric Methods to Evaluate Measurement and Algorithmic Bias in Automated Scoring. *Journal of Educational Measurement* 59, 3 (2022), 338–361. doi:10.1111/jedm.12335

[189] Mackenzie Jorgensen, Hannah Richert, Elizabeth Black, Natalia Criado, and Jose Such. 2023. Not so Fair: The Impact of Presumably Fair Machine Learning Models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (Aies '23)*. Association for Computing Machinery, New York, NY, USA, 297–311. doi:10.1145/3600211.3604699

[190] Guusje Juijn, Niya Stoimenova, João Reis, and Dong Nguyen. 2023. Perceived Algorithmic Fairness Using Organizational Justice Theory: An Empirical Case Study on Algorithmic Hiring. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (Aies '23)*. Association for Computing Machinery, New York, NY, USA, 775–785. doi:10.1145/3600211.3604677

[191] Md Alamgir Kabir, Mobyen Uddin Ahmed, Shahina Begum, Shaibal Barua, and Md Rakibul Islam. 2024. Balancing Fairness: Unveiling the Potential of SMOTE-driven Oversampling in AI Model Enhancement. In *Proceedings of the 2024 9th International Conference on Machine Learning Technologies (Icmlt '24)*. Association for Computing Machinery, New York, NY, USA, 21–29. doi:10.1145/3674029.3674034

[192] N. Kallus, X. Mao, and A. Zhou. 2022. Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination. *Management Science* 68, 3 (2022), 1959–1981. doi:10.1287/mnsc.2020.3850

[193] Parameswaran Kamalaruban, Yulu Pi, Stuart Burrell, Eleanor Drage, Piotr Skalski, Jason Wong, and David Sutton. 2024. Evaluating Fairness in Transaction Fraud Models: Fairness Metrics, Bias Audits, and Challenges. In *Proceedings of the 5th ACM International Conference on AI in Finance (Icaif '24)*. Association for Computing Machinery, New York, NY, USA, 555–563. doi:10.1145/3677052.3698666

[194] S. Karasaki, R. Morello-Frosch, and D. Callaway. 2024. Machine Learning for Environmental Justice: Dissecting an Algorithmic Approach to Predict Drinking Water Quality in California. *Science of the Total Environment* 951 (2024), 1–10. doi:10.1016/j.scitotenv.2024.175730

[195] Marzieh Karimi-Haghighi and Carlos Castillo. 2021. Enhancing a Recidivism Prediction Tool with Machine Learning: Effectiveness and Algorithmic Fairness. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law (Icail '21)*. Association for Computing Machinery, New York, NY, USA, 210–214. doi:10.1145/3462757.3466150

[196] S. Khor, E.C. Haupt, E.E. Hahn, L.J.L. Lyons, V. Shankaran, and A. Bansal. 2023. Racial and Ethnic Bias in Risk Prediction Models for Colorectal Cancer Recurrence When Race and Ethnicity Are Omitted as Predictors. *JAMA Network Open* 6, 6 (2023), E2318495. doi:10.1001/jamanetworkopen.2023.18495

[197] Jin-Young Kim and Sung-Bae Cho. 2022. An Information Theoretic Approach to Reducing Algorithmic Bias for Machine Learning. *Neurocomputing* 500 (Aug. 2022), 26–38. doi:10.1016/j.neucom.2021.09.081

[198] Michael P. Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-box Post-Processing for Fairness in Classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (Aies '19)*. Association for Computing Machinery, New York, NY, USA, 247–254. doi:10.1145/3306618.3314287

[199] S. Kim, S. Kim, K. Kim, and Y. Kim. 2023. Lq Regularization for Fair Artificial Intelligence Robust to Covariate Shift. *Statistical Analysis and Data Mining* 16, 3 (2023), 237–256. doi:10.1002/sam.11616

[200] Y. Kim and H. Jeong. 2022. Learning Fair Prediction Models with an Imputed Sensitive Variable: Empirical Studies. *Communications for Statistical Applications and Methods* 29, 2 (2022), 251–261. doi:10.29220/CSAM.2022.29.2.251

[201] M. Kisten and M. Khosa. 2024. Enhancing Fairness in Credit Assessment: Mitigation Strategies and Implementation. *IEEE Access* 12 (2024), 177277–177284. doi:10.1109/ACCESS.2024.3505836

[202] D. Kiyasseh, J. Laca, T.F. Haque, M. Otiato, B.J. Miles, C. Wagner, D.A. Donoho, Q.-D. Trinh, A. Anandkumar, and A.J. Hung. 2023. Human Visual Explanations Mitigate Bias in AI-based Assessment of Surgeon Skills. *npj Digital Medicine* 6, 1 (2023), 1–12. doi:10.1038/s41746-023-00766-2

[203] M. Klingenberg, D. Stark, F. Eitel, C. Budding, M. Habes, and K. Ritter. 2023. Higher Performance for Women than Men in MRI-based Alzheimer's Disease Detection. *Alzheimer's Research and Therapy* 15, 1 (2023), 1–13. doi:10.1186/s13195-023-01225-6

[204] I. Kong, K. Kim, and Y. Kim. 2025. Fair Representation Learning for Continuous Sensitive Attributes Using Expectation of Integral Probability Metrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47, 5 (2025), 3784–3795. doi:10.1109/TPAMI.2025.3538915

[205] M. Kuppler. 2022. Predicting the Future Impact of Computer Science Researchers: Is There a Gender Bias? *Scientometrics* 127, 11 (2022), 6695–6732. doi:10.1007/s11192-022-04337-2

[206] Selim Kuzucu, Jiaee Cheong, Hatice Gunes, and Sinan Kalkan. 2024. Uncertainty as a Fairness Measure. *Journal of Artificial Intelligence Research* 81 (Sept. 2024), 307–335. doi:10.1613/jair.1.16041

[207] G.H. Kwak, H.A. Kamdar, M.J. Douglas, H. Hu, S.E. Ack, I.A. Lissak, A.E. Williams, N. Yechoor, and E.S. Rosenthal. 2024. Social Determinants of Health and Limitation of Life-Sustaining Therapy in Neurocritical Care: A CHoRUS Pilot Project. *Neurocritical Care* 41, 3 (2024), 866–879. doi:10.1007/s12028-024-02007-0

[208] A. Langenberg, S.-C. Ma, T. Ermakova, and B. Fabian. 2023. Formal Group Fairness and Accuracy in Automated Decision Making. *Mathematics* 11, 8 (2023), 1–25. doi:10.3390/math11081771

[209] Daphne Lenders and Toon Calders. 2023. Real-Life Performance of Fairness Interventions - Introducing a New Benchmarking Dataset for Fair ML. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing (Sac '23)*. Association for Computing Machinery, New York, NY, USA, 350–357. doi:10.1145/3555776.3577634

[210] C. Li, S. Ding, N. Zou, X. Hu, X. Jiang, and K. Zhang. 2023. Multi-Task Learning with Dynamic Re-Weighting to Achieve Fairness in Healthcare Predictive Modeling. *Journal of Biomedical Informatics* 143 (2023), 1–11. doi:10.1016/j.jbi.2023.104399

[211] Chenglu Li, Wanli Xing, and Walter Leite. 2024. Using Fair AI to Predict Students' Math Learning Outcomes in an Online Platform. *Interactive Learning Environments* 32, 3 (May 2024), 1117–1136. doi:10.1080/10494820.2022.2115076

[212] Fuchen Li, Patrick Wu, Henry H. Ong, Josh F. Peterson, Wei-Qi Wei, and Juan Zhao. 2023. Evaluating and Mitigating Bias in Machine Learning Models for Cardiovascular Disease Prediction. *Journal of Biomedical Informatics* 138 (Feb. 2023), N.PAG–N.PAG. doi:10.1016/j.jbi.2023.104294

[213] Nianyun Li, Naman Goel, and Elliott Ash. 2022. Data-Centric Factors in Algorithmic Fairness. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (Aies '22)*. Association for Computing Machinery, New York, NY, USA, 396–410. doi:10.1145/3514094.3534147

[214] Xinyue Li, Zhenpeng Chen, Jie M. Zhang, Federica Sarro, Ying Zhang, and Xuanzhe Liu. 2025. Bias behind the Wheel: Fairness Testing of Autonomous Driving Systems. *ACM Transactions on Software Engineering and Methodology* 34, 3, Article 82 (Feb. 2025), 24 pages. doi:10.1145/3702989

[215] Y. Li, H. Wang, and Y. Luo. 2022. Improving Fairness in the Prediction of Heart Failure Length of Stay and Mortality by Integrating Social Determinants of Health. *Circulation: Heart Failure* 15, 11 (2022), E009473. doi:10.1161/CIRCHEARTFAILURE.122.009473

[216] Y. Liang, H. Chao, J. Zhang, G. Wang, and P. Yan. 2024. Unbiasing Fairness Evaluation of Radiology AI Model. *Meta-Radiology* 2, 3 (2024), 1–7. doi:10.1016/j.metrad.2024.100084

[217] Maria Liatsikou, Sofia Yfantidou, and Athena Vakali. 2024. Assessing Bias in Self-Supervised and Supervised Models for Temporal Data. In *Proceedings of the 13th Hellenic Conference on Artificial Intelligence (Setn '24)*. Association for Computing Machinery, New York, NY, USA, Article 1, 7 pages. doi:10.1145/3688671.3688753

[218] M. Lin, T. Li, Z. Sun, G. Holste, Y. Ding, F. Wang, G. Shih, and Y. Peng. 2024. Improving Fairness of Automated Chest Radiograph Diagnosis by Contrastive Learning. *Radiology: Artificial Intelligence* 6, 5 (2024), 1–10. doi:10.1148/ryai.230342

[219] David Liu, Zohair Shafi, William Fleisher, Tina Eliassi-Rad, and Scott Alfeld. 2021. RAWLSNET: Altering Bayesian Networks to Encode Rawlsian Fair Equality of Opportunity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (Aies '21)*. Association for Computing Machinery, New York, NY, USA, 745–755. doi:10.1145/3461702.3462618

[220] M. Liu, Y. Ning, Y. Ke, Y. Shang, B. Chakraborty, M.E.H. Ong, R. Vaughan, and N. Liu. 2024. FAIM: Fairness-aware Interpretable Modeling for Trustworthy Machine Learning in Healthcare. *Patterns* 5, 10 (2024), 14. doi:10.1016/j.patter.2024.101059

[221] W. Liu, X. Wang, H. Zheng, B. Jin, X. Wang, and H. Zha. 2022. Mitigating Disparate Impact on Model Accuracy in Differentially Private Learning. *Information Sciences* 616 (2022), 108–126. doi:10.1016/j.ins.2022.10.019

[222] Zhining Liu, Ruizhong Qiu, Zhichen Zeng, Yada Zhu, Hendrik Hamann, and Hanghang Tong. 2024. AIM: Attributing, Interpreting, Mitigating Data Unfairness. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Kdd '24)*. Association for Computing Machinery, New York, NY, USA, 2014–2025. doi:10.1145/3637528.3671797

[223] Z. Liu, X. Zhang, and B. Jiang. 2023. Active Learning with Fairness-Aware Clustering for Fair Classification Considering Multiple Sensitive Attributes. *Information Sciences* 647 (2023), 1–16. doi:10.1016/j.ins.2023.119521

[224] R.B. Loureiro, T.P. Pagano, F.V.N. Lisboa, L.F.S. Nascimento, E.L.S. Oliveira, I. Winkler, and E.G.S. Nascimento. 2025. Correlation-Based Methods for Representative Fairness Metric Selection: An Empirical Study on Efficiency and Caveats in Model Evaluation. *Expert Systems with Applications* 268 (2025), 1–20. doi:10.1016/j.eswa.2024.126344

[225] Kristian Lum, Yunfeng Zhang, and Amanda Bower. 2022. De-Biasing "Bias" Measurement. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 379–389. doi:10.1145/3531146.3533105

[226] Jing Ma, Ruocheng Guo, Aidong Zhang, and Jundong Li. 2023. Learning for Counterfactual Fairness from Observational Data. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Kdd '23)*. Association for Computing Machinery, New York, NY, USA, 1620–1630. doi:10.1145/3580305.3599408

[227] P. Majumdar, S. Chhabra, R. Singh, and M. Vatsa. 2021. Subgroup Invariant Perturbation for Unbiased Pre-Trained Model Prediction. *Frontiers in Big Data* 3 (2021), 1–16. doi:10.3389/fdata.2020.590296

[228] Suvodeep Majumder, Joymallya Chakraborty, Gina R. Bai, Kathryn T. Stolee, and Tim Menzies. 2023. Fair Enough: Searching for Sufficient Measures of Fairness. *ACM Transactions on Software Engineering and Methodology* 32, 6, Article 134 (Sept. 2023), 21 pages. doi:10.1145/3585006

[229] Marta Marchiori Manerba and Riccardo Guidotti. 2022. Investigating Debiasing Effects on Classification and Explainability. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (Aies '22)*. Association for Computing Machinery, New York, NY, USA, 468–478. doi:10.1145/3514094.3534170

[230] Mariachiara Mecati, Antonio Vetrò, and Marco Torchiano. 2022. Detecting Risk of Biased Output with Balance Measures. *ACM Journal of Data and Information Quality* 14, 4, Article 25 (Nov. 2022), 7 pages. doi:10.1145/3530787

[231] Esther L. Meerwijk, Duncan C. McElfresh, Susana Martins, and Suzanne R. Tamang. 2024. Evaluating Accuracy and Fairness of Clinical Decision Support Algorithms When Health Care Resources Are Limited. *Journal of Biomedical Informatics* 156 (Aug. 2024), N.PAG–N.PAG. doi:10.1016/j.jbi.2024.104664

[232] C. Meng, L. Trinh, N. Xu, J. Enouen, and Y. Liu. 2022. Interpretability and Fairness Evaluation of Deep Learning Models on MIMIC-IV Dataset. *Scientific Reports* 12, 1 (2022), 1–28. doi:10.1038/s41598-022-11012-2

[233] C. Mennella, M. Esposito, G. De Pietro, and U. Maniscalco. 2024. Promoting Fairness in Activity Recognition Algorithms for Patient's Monitoring and Evaluation Systems in Healthcare. *Computers in Biology and Medicine* 179 (2024), 1–14. doi:10.1016/j.compbiomed.2024.108826

[234] Vishwali Mhasawade and Rumi Chunara. 2021. Causal Multi-Level Fairness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (Aies '21)*. Association for Computing Machinery, New York, NY, USA, 784–794. doi:10.1145/3461702.3462587

[235] Marius Miron, Songül Tolan, Emilia Gómez, and Carlos Castillo. 2021. Evaluating Causes of Algorithmic Bias in Juvenile Criminal Recidivism. *Artificial Intelligence & Law* 29, 2 (June 2021), 111–147. doi:10.1007/s10506-020-09268-y

[236] A. Miroshnikov, K. Kotsiopoulos, R. Franks, and A. Ravi Kannan. 2022. Wasserstein-Based Fairness Interpretability Framework for Machine Learning Models. *Machine Learning* 111, 9 (2022), 3307–3357. doi:10.1007/s10994-022-06213-9

[237] J. Misztal-Radecka and B. Indurkhya. 2023. A Bias Detection Tree Approach for Detecting Disparities in a Recommendation Model's Errors. *User Modeling and User-Adapted Interaction* 33, 1 (2023), 43–79. doi:10.1007/s11257-022-09334-x

[238] Kiarash Mohammadi, Aishwarya Sivaraman, and Golnoosh Farnadi. 2023. FETA: Fairness Enforced Verifying, Training, and Predicting Algorithms for Neural Networks. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (Eaamo '23)*. Association for Computing Machinery, New York, NY, USA, Article 23, 13 pages. doi:10.1145/3617694.3623243

[239] E. Montana, D.S. Nagin, R. Neil, and R.J. Sampson. 2023. Cohort Bias in Predictive Risk Assessments of Future Criminal Justice System Involvement. *Proceedings of the National Academy of Sciences of the United States of America* 120, 23 (2023), 1–9. doi:10.1073/pnas.2301990120

[240] M. Moriah, F. Vermet, and A. Charpentier. 2024. Measuring and Mitigating Biases in Motor Insurance Pricing. *European Actuarial Journal* 14, 3 (2024), 833–869. doi:10.1007/s13385-024-00390-8

[241] P. Mosteiro, J. Kuiper, J. Masthoff, F. Scheepers, and M. Spruit. 2022. Bias Discovery in Machine Learning Models for Mental Health. *Information (Switzerland)* 13, 5 (2022), 1–15. doi:10.3390/info13050237

[242] Carlos Mougan, Jose Manuel Alvarez, Salvatore Ruggieri, and Steffen Staab. 2023. Fairness Implications of Encoding Protected Categorical Attributes. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (Aies '23)*. Association for Computing Machinery, New York, NY, USA, 454–465. doi:10.1145/3600211.3604657

[243] W.S. Murch, S. Kairouz, and M. French. 2024. Comparing 'Fair' Machine Learning Models for Detecting at-Risk Online Gamblers. *International Gambling Studies* 15, 1 (2024), 84–106. doi:10.1080/14459795.2024.2412051

[244] Ece Çiğdem Mutlu, Niloofar Yousefi, and Ozlem Ozmen Garibay. 2022. Contrastive Counterfactual Fairness in Algorithmic Decision-Making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (Aies '22)*. Association for Computing Machinery, New York, NY, USA, 499–507. doi:10.1145/3514094.3534143

[245] R. Nagpal, A. Khan, M. Borkar, and A. Gupta. 2024. A Multi-Objective Framework for Balancing Fairness and Accuracy in Debiasing Machine Learning Models. *Machine Learning and Knowledge Extraction* 6, 3 (2024), 2130–2148. doi:10.3390/make6030105

[246] E.D.A.L. Neto, J. Bailiss, A. Finke, J. Miller, and G. Cosma. 2024. Identifying Early Help Referrals for Local Authorities with Machine Learning and Bias Analysis. *Journal of Computational Social Science* 7, 1 (2024), 385–403. doi:10.1007/s42001-023-00242-7

[247] N. Nezami, P. Haghighat, D. Gándara, and H. Anahideh. 2024. Assessing Disparities in Predictive Modeling Outcomes for College Student Success: The Impact of Imputation Techniques on Model Performance and Fairness. *Education Sciences* 14, 2 (2024), 1–24. doi:10.3390/educsci14020136

[248] Giang Nguyen, Sumon Biswas, and Hridesh Rajan. 2023. Fix Fairness, Don't Ruin Accuracy: Performance Aware Fairness Repair Using AutoML. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (Esec/Fse 2023)*. Association for Computing Machinery, New York, NY, USA, 502–514. doi:10.1145/3611643.3616257

[249] David Nigenda, Zohar Karnin, Muhammad Bilal Zafar, Raghu Ramesha, Alan Tan, Michele Donini, and Krishnaram Kenthapadi. 2022. Amazon SageMaker Model Monitor: A System for Real-Time Insights into Deployed Machine Learning Models. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Kdd '22)*. Association for Computing Machinery, New York, NY, USA, 3671–3681. doi:10.1145/3534678.3539145

[250] Luca Oneto, Michele Doninini, Amon Elders, and Massimiliano Pontil. 2019. Taking Advantage of Multitask Learning for Fair Classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (Aies '19)*. Association for Computing Machinery, New York, NY, USA, 227–237. doi:10.1145/3306618.3314255

[251] M. Openja, G. Laberge, and F. Khomh. 2024. Detection and Evaluation of Bias-Inducing Features in Machine Learning. *Empirical Software Engineering* 29, 1 (2024), 1–71. doi:10.1007/s10664-023-10409-5

[252] T.P. Pagano, R.B. Loureiro, F.V.N. Lisboa, G.O.R. Cruz, R.M. Peixoto, G.A.D.S. Guimarães, E.L.S. Oliveira, I. Winkler, and E.G.S. Nascimento. 2023. Context-Based Patterns in Machine Learning Bias and Fairness Metrics: A Sensitive Attributes-Based Approach. *Big Data and Cognitive Computing* 7, 1 (2023), 1–18. doi:10.3390/bdcc7010027

[253] Cecilia Panigutti, Alan Perotti, André Panisson, Paolo Bajardi, and Dino Pedreschi. 2021. FairLens: Auditing Black-Box Clinical Decision Support Systems. *Information Processing & Management* 58, 5 (Sept. 2021), N.PAG–N.PAG. doi:10.1016/j.ipm.2021.102657

[254] J. Park, R. Arunachalam, V. Silenzio, and V.K. Singh. 2022. Fairness in Mobile Phone–Based Mental Health Assessment Algorithms: Exploratory Study. *JMIR Formative Research* 6, 6 (2022), 1–11. doi:10.2196/34366

[255] Y. Park, J. Hu, M. Singh, I. Sylla, I. Dankwa-Mullan, E. Koski, and A.K. Das. 2021. Comparison of Methods to Reduce Bias from Clinical Prediction Models of Postpartum Depression. *JAMA Network Open* 4, 4 (2021), 1–15. doi:10.1001/jamanetworkopen.2021.3909

[256] P. Patil and K. Purcell. 2022. Decorrelation-Based Deep Learning for Bias Mitigation. *Future Internet* 14, 4 (2022), 1–14. doi:10.3390/fi14040110

[257] William Paul, Armin Hadzic, Neil Joshi, Fady Alajaji, and Philippe Burlina. 2022. TARA: Training and Representation Alteration for AI Fairness and Domain Generalization. *Neural Computation* 34, 3 (March 2022), 716–753. doi:10.1162/neco_a_01468

[258] Ángel Pavón Pérez, Miriam Fernandez, Hasan Al-Madfai, Grégoire Burel, and Harith Alani. 2023. Tracking Machine Learning Bias Creep in Traditional and Online Lending Systems with Covariance Analysis. In *Proceedings of the 15th ACM Web Science Conference 2023 (WebSci '23)*. Association for Computing Machinery, New York, NY, USA, 184–195. doi:10.1145/3578503.3583605

[259] G.D. Pelegrina and L.T. Duarte. 2024. A Novel Approach for Fair Principal Component Analysis Based on Eigendecomposition. *IEEE Transactions on Artificial Intelligence* 5, 3 (2024), 1195–1206. doi:10.1109/TAI.2023.3298291

[260] K. Peng, J. Chakraborty, and T. Menzies. 2023. FairMask: Better Fairness via Model-Based Rebalancing of Protected Attributes. *IEEE Transactions on Software Engineering* 49, 4 (2023), 2426–2439. doi:10.1109/TSE.2022.3220713

[261] A. Perera, A. Aleti, C. Tantithamthavorn, J. Jiarpakdee, B. Turhan, L. Kuhn, and K. Walker. 2022. Search-Based Fairness Testing for Regression-Based Machine Learning Systems. *Empirical Software Engineering* 27, 3 (2022), 1–36. doi:10.1007/s10664-022-10116-7

[262] Valerio Perrone, Michele Donini, Muhammad Bilal Zafar, Robin Schmucker, Krishnaram Kenthapadi, and Cédric Archambeau. 2021. Fair Bayesian Optimization. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (Aies '21)*. Association for Computing Machinery, New York, NY, USA, 854–863. doi:10.1145/3461702.3462629

[263] Dana Pessach and Erez Shmueli. 2021. Improving Fairness of Artificial Intelligence Algorithms in Privileged-Group Selection Bias Data Settings. *Expert Systems with Applications* 185 (Dec. 2021), N.PAG–N.PAG. doi:10.1016/j.eswa.2021.115667

[264] Eike Petersen, Melanie Ganz, Sune Holm, and Aasa Feragen. 2023. On (Assessing) the Fairness of Risk Score Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 817–829. doi:10.1145/3593013.3594045

[265] Andrija Petrović, Mladen Nikolić, Miloš Jovanović, Miloš Bijanić, and Boris Delibašić. 2021. Fair Classification via Monte Carlo Policy Gradient Method. *Engineering Applications of Artificial Intelligence* 104 (Sept. 2021), N.PAG–N.PAG. doi:10.1016/j.engappai.2021.104398

[266] Andrija Petrović, Mladen Nikolić, Sandro Radovanović, Boris Delibašić, and Miloš Jovanović. 2022. FAIR: Fair Adversarial Instance Re-Weighting. *Neurocomputing* 476 (March 2022), 14–37. doi:10.1016/j.neucom.2021.12.082

[267] F. Pfisterer, S. Wei, S. Vollmer, M. Lang, and B. Bischl. 2023. Fairness Audits and Debiasing Using Mlr3fairness. *R Journal* 15, 1 (2023), 234–253. doi:10.32614/RJ-2023-034

[268] N. Pham, M.K. Do, T.V. Dai, P.N. Hung, and A. Nguyen-Duc. 2025. FAIREDU: A Multiple Regression-Based Method for Enhancing Fairness in Machine Learning Models for Educational Applications. *Expert Systems with Applications* 269 (2025), 1–18. doi:10.1016/j.eswa.2024.126219

[269] Thomas Pinkava, Jack McFarland, and Afra Mashhadi. 2025. A Model- and Data-Agnostic Debiasing System for Achieving Equalized Odds. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (Aies '24)*. AAAI Press, San Jose, California, USA, 1123–1131.

[270] G. Popoola and J. Sheppard. 2024. Investigating and Mitigating the Performance–Fairness Tradeoff via Protected-Category Sampling. *Electronics (Switzerland)* 13, 15 (2024), 24. doi:10.3390/electronics13153024

[271] L. Prior and G. Leckie. 2024. Student Intersectional Sociodemographic and School Variation in GCSE Final Grades in England Following Covid-19 Examination Cancellations. *Oxford Review of Education* (2024), 23 pages. doi:10.1080/03054985.2024.2407624

[272] E. Puyol-Antón, B. Ruijsink, J. Mariscal Harana, S.K. Piechnik, S. Neubauer, S.E. Petersen, R. Razavi, P. Chowienczyk, and A.P. King. 2022. Fairness in Cardiac Magnetic Resonance Imaging: Assessing Sex and Racial Bias in Deep Learning-Based Segmentation. *Frontiers in Cardiovascular Medicine* 9 (2022), 1–10. doi:10.3389/fcvm.2022.859310

[273] H. Qiu, R. Feng, R. Hu, X. Yang, S. Lin, Q. Tao, and Y. Yang. 2023. Learning Fair Representations via an Adversarial Framework. *AI Open* 4 (2023), 91–97. doi:10.1016/j.aiopen.2023.08.003

[274] Tangkun Quan, Fei Zhu, Quan Liu, and Fanzhang Li. 2023. Learning Fair Representations for Accuracy Parity. *Engineering Applications of Artificial Intelligence* 119 (March 2023), N.PAG–N.PAG. doi:10.1016/j.engappai.2023.105819

[275] E.M. Queiroga, M.F. Batista Machado, V.R. Paragarino, T.T. Primo, and C. Cechinel. 2022. Early Prediction of At-Risk Students in Secondary Education: A Countrywide K-12 Learning Analytics Initiative in Uruguay. *Information (Switzerland)* 13, 9 (2022), 1–25. doi:10.3390/info13090401

[276] Bilal Qureshi, Faisal Kamiran, Asim Karim, Salvatore Ruggieri, and Dino Pedreschi. 2020. Causal Inference for Social Discrimination Reasoning. *Journal of Intelligent Information Systems* 54, 2 (April 2020), 425–437. doi:10.1007/s10844-019-00580-x

[277] S. Radovanović, A. Petrović, B. Delibašić, and M. Suknović. 2023. A Fair Classifier Chain for Multi-Label Bank Marketing Strategy Classification. *International Transactions in Operational Research* 30, 3 (2023), 1320–1339. doi:10.1111/itor.13059

[278] G. Raftopoulos, G. Davrazos, and S. Kotsiantis. 2024. Fair and Transparent Student Admission Prediction Using Machine Learning Models. *Algorithms* 17, 12 (2024), 1–19. doi:10.3390/a17120572

[279] S. Rao and T. Zhao. 2025. Ethical AI in HR: A Case Study of Tech Hiring. *Journal of Computer Information Systems* (2025), 19 pages. doi:10.1080/08874417.2024.2446954

[280] R. Ravindranath, J.D. Stein, T. Hernandez-Boussard, A.C. Fisher, S.Y. Wang, S. Amin, P.A. Edwards, D. Srikumaran, F. Woreta, J.S. Schultz, A. Shrivastava, B. Ahmad, P. Bryar, D. French, B.L. Vanderbeek, S. Pershing, A.M. Lynch, J.L. Patnaik, S. Munir, W. Munir, L. DeLott, B.C. Stagg, B. Wirostko, B. McMillian, A. Sheybani, S. Sarrapour, K. Nwanyanwu, M. Deiner, C. Sun, H.R. Feldman, and R. Ramachandran. 2025. The Impact of Race, Ethnicity, and Sex on Fairness in Artificial Intelligence for Glaucoma Prediction Models. *Ophthalmology Science* 5, 1 (2025), 1–10. doi:10.1016/j.xops.2024.100596

[281] Brianna Richardson, Prasanna Sattigeri, Dennis Wei, Karthikeyan Natesan Ramamurthy, Kush Varshney, Amit Dhurandhar, and Juan E. Gilbert. 2023. Add-Remove-or-Relabel: Practitioner-friendly Bias Mitigation via Influential Fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 736–752. doi:10.1145/3593013.3594039

[282] Jake Robertson, Catherine Stinson, and Ting Hu. 2022. A Bio-Inspired Framework for Machine Bias Interpretation. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (Aies '22)*. Association for Computing Machinery, New York, NY, USA, 588–598. doi:10.1145/3514094.3534126

[283] E. Röösli, S. Bozkurt, and T. Hernandez-Boussard. 2022. Peeking into a Black Box, the Fairness and Generalizability of a MIMIC-III Benchmarking Model. *Scientific Data* 9, 1 (2022), 1–13. doi:10.1038/s41597-021-01110-7

[284] Alveiro Alonso Rosado Gómez, Maritza Liliana Calderón Benavides, and Oscar Espinosa. 2024. Data Preprocessing to Improve Fairness in Machine Learning Models: An Application to the Reintegration Process of Demobilized Members of Armed Groups in Colombia. *Applied Soft Computing* 152 (Feb. 2024), N.PAG–N.PAG. doi:10.1016/j.asoc.2023.111193

[285] Matteo Ruggeri, Alice Dethise, and Marco Canini. 2023. On Detecting Biased Predictions with Post-Hoc Explanation Methods. In *Proceedings of the 2023 on Explainable and Safety Bounded, Fidelitous, Machine Learning for Networking (Safe '23)*. Association for Computing Machinery, New York, NY, USA, 17–23. doi:10.1145/3630050.3630179

[286] D. Şahin, L. Kambeitz-Ilankovic, S. Wood, D. Dwyer, R. Upthegrove, R. Salokangas, S. Borgwardt, P. Brambilla, E. Meisenzahl, S. Ruhrmann, F. Schultze-Lutter, R. Lencer, A. Bertolino, C. Pantelis, N. Koutsouleris, and J. Kambeitz. 2024. Algorithmic Fairness in Precision Psychiatry: Analysis of Prediction Models in Individuals at Clinical High Risk for Psychosis. *British Journal of Psychiatry* 224, 2 (2024), 55–65. doi:10.1192/bjp.2023.141

[287] Z. Salahuddin, Y. Chen, X. Zhong, H.C. Woodruff, N.M. Rad, S.A. Mali, and P. Lambin. 2023. From Head and Neck Tumour and Lymph Node Segmentation to Survival Prediction on PET/CT: An End-to-End Framework Featuring Uncertainty, Fairness, and Multi-Region Multi-Modal Radiomics. *Cancers* 15, 7 (2023), 1–21. doi:10.3390/cancers15071932

[288] T. Salazar, M.S. Santos, H. Araujo, and P.H. Abreu. 2021. FAWOS: Fairness-Aware Oversampling Algorithm Based on Distributions of Sensitive Attributes. *IEEE Access* 9 (2021), 81370–81379. doi:10.1109/ACCESS.2021.3084121

[289] Jakob Schoeffer, Niklas Kuehl, and Isabel Valera. 2021. A Ranking Approach to Fair Classification. In *Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies (Compass '21)*. Association for Computing Machinery, New York, NY, USA, 115–125. doi:10.1145/3460112.3471950

[290] H.S. Schuch, M. Furtado, G.F.D.S. Silva, I. Kawachi, A.D.P. Chiavegatto Filho, and H.W. Elani. 2023. Fairness of Machine Learning Algorithms for Predicting Foregone Preventive Dental Care for Adults. *JAMA Network Open* 6, 11 (2023), E2341625. doi:10.1001/jamanetworkopen.2023.41625

[291] Ignacio Serna, Aythami Morales, Julian Fierrez, and Nick Obradovich. 2022. Sensitive Loss: Improving Accuracy and Fairness of Face Representations with Discrimination-Aware Deep Learning. *Artificial Intelligence* 305 (April 2022), N.PAG–N.PAG. doi:10.1016/j.artint.2022.103682

[292] Lele Sha, Dragan Gašević, and Guanliang Chen. 2023. Lessons from Debiasing Data for Fair and Accurate Predictive Modeling in Education. *Expert Systems with Applications* 228 (Oct. 2023), N.PAG–N.PAG. doi:10.1016/j.eswa.2023.120323

[293] L. Sha, M. Rakovic, A. Das, D. Gasevic, and G. Chen. 2022. Leveraging Class Balancing Techniques to Alleviate Algorithmic Bias for Predictive Tasks in Education. *IEEE Transactions on Learning Technologies* 15, 4 (2022), 481–492. doi:10.1109/TLT.2022.3196278

[294] Shubham Sharma, Alan H. Gee, David Paydarfar, and Joydeep Ghosh. 2021. FaiR-n: Fair and Robust Neural Networks for Structured Data. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (Aies '21)*. Association for Computing Machinery, New York, NY, USA, 946–955. doi:10.1145/3461702.3462559

[295] Arumoy Shome, Luis Cruz, and Arie Van Deursen. 2024. Data vs. Model Machine Learning Fairness Testing: An Empirical Study. In *Proceedings of the 5th IEEE/ACM International Workshop on Deep Learning for Testing and Testing for Deep Learning (DeepTest '24)*. Association for Computing Machinery, New York, NY, USA, 1–8. doi:10.1145/3643786.3648022

[296] I.A. Siddiqui, N. Littlefield, L.A. Carlson, M. Gong, A. Chhabra, Z. Menezes, G.M. Mastorakos, S.M. Thakar, M. Abedian, I. Lohse, K.R. Weiss, J.F. Plate, H. Moradi, S. Amirian, and A.P. Tafti. 2024. Fair AI-powered Orthopedic Image Segmentation: Addressing Bias and Promoting Equitable Healthcare. *Scientific Reports* 14, 1 (2024), 1–16. doi:10.1038/s41598-024-66873-6

[297] Jan Simson, Florian Pfisterer, and Christoph Kern. 2024. One Model Many Scores: Using Multiverse Analysis to Prevent Fairness Hacking and Evaluate the Influence of Model Design Decisions. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1305–1320. doi:10.1145/3630106.3658974

[298] A. Singh, J. Singh, A. Khan, and A. Gupta. 2022. Developing a Novel Fair-Loan Classifier through a Multi-Sensitive Debiasing Pipeline: DualFair. *Machine Learning and Knowledge Extraction* 4, 1 (2022), 240–253. doi:10.3390/make4010011

[299] S. Sivarajkumar, Y. Huang, and Y. Wang. 2023. Fair Patient Model: Mitigating Bias in the Patient Representation Learned from the Electronic Health Records. *Journal of Biomedical Informatics* 148 (2023), 1–25. doi:10.1016/j.jbi.2023.104544

[300] Gizem Sogancioglu, Pablo Mosteiro, Albert Ali Salah, Floortje Scheepers, and Heysem Kaya. 2025. Fairness in AI-based Mental Health: Clinician Perspectives and Bias Mitigation. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (Aies '24)*. AAAI Press, San Jose, California, USA, 1390–1400. doi:10.1609/aies.v7i1.31732

[301] David Solans Noguero, Diana Ramírez-Cifuentes, Esteban Andrés Ríssola, and Ana Freire. 2023. Gender Bias When Using Artificial Intelligence to Assess Anorexia Nervosa on Social Media: Data-Driven Study. *Journal of Medical Internet Research* 25, 1 (Jan. 2023), 1–19. doi:10.2196/45184

[302] Serban Stan and Mohammad Rostami. 2024. Preserving Fairness in AI under Domain Shift. *Journal of Artificial Intelligence Research* 81 (Sept. 2024), 907–934. doi:10.1613/jair.1.16694

[303] Isabel Straw, Geraint Rees, and Parashkev Nachev. 2024. Sex-Based Performance Disparities in Machine Learning Algorithms for Cardiac Disease Prediction: Exploratory Study. *Journal of Medical Internet Research* 26 (Jan. 2024), 1–18. doi:10.2196/46936

[304] I. Straw and H. Wu. 2022. Investigating for Bias in Healthcare Algorithms: A Sex-Stratified Analysis of Supervised Machine Learning Models in Liver Disease Prediction. *BMJ Health and Care Informatics* 29, 1 (2022), 1–8. doi:10.1136/bmjhci-2021-100457

[305] M.A. Sufian, L. Alsadder, W. Hamzi, S. Zaman, A.S.M.S. Sagar, and B. Hamzi. 2024. Mitigating Algorithmic Bias in AI-Driven Cardiovascular Imaging for Fairer Diagnostics. *Diagnostics* 14, 23 (2024), 1–55. doi:10.3390/diagnostics14232675

[306] Y. Suk and K.T. Han. 2024. Evaluating Intersectional Fairness in Algorithmic Decision Making Using Intersectional Differential Algorithmic Functioning. *Journal of Educational and Behavioral Statistics* 50, 5 (2024), 1–24. doi:10.3102/10769986241269820

[307] Y. Suk and K.T. Han. 2024. A Psychometric Framework for Evaluating Fairness in Algorithmic Decision Making: Differential Algorithmic Functioning. *Journal of Educational and Behavioral Statistics* 49, 2 (2024), 151–172. doi:10.3102/10769986231171711

[308] W. Tan, Q. Wei, Z. Xing, H. Fu, H. Kong, Y. Lu, B. Yan, and C. Zhao. 2024. Fairer AI in Ophthalmology via Implicit Fairness Learning for Mitigating Sexism and Ageism. *Nature Communications* 15, 1 (2024), 1–13. doi:10.1038/s41467-024-48972-0

[309] Z. Tang, T. Lu, and T. Li. 2024. Metric-Independent Mitigation of Unpredefined Bias in Machine Classification. *Intelligent Computing* 3 (2024), 1–15. doi:10.34133/icomputing.0083

[310] Bahar Taskesen, Jose Blanchet, Daniel Kuhn, and Viet Anh Nguyen. 2021. A Statistical Test for Probabilistic Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 648–665. doi:10.1145/3442188.3445927

[311] L. Ter-Minassian, N. Viani, A. Wickersham, L. Cross, R. Stewart, S. Velupillai, and J. Downs. 2022. Assessing Machine Learning for Fair Prediction of ADHD in School Pupils Using a Retrospective Cohort Study of Linked Education and Healthcare Data. *BMJ Open* 12, 12 (2022), 1–9. doi:10.1136/bmjopen-2021-058058

[312] B. Theodorou, B. Danek, V. Tummala, S.P. Kumar, B. Malin, and J. Sun. 2025. Improving Medical Machine Learning Models with Generative Balancing for Equity and Excellence. *npj Digital Medicine* 8, 1 (2025), 1–11. doi:10.1038/s41746-025-01438-z

[313] Hale M Thompson, Brihat Sharma, Sameer Bhalla, Randy Boley, Connor McCluskey, Dmitriy Dligach, Matthew M Churpek, Niranjan S Karnik, and Majid Afshar. 2021. Bias and Fairness Assessment of a Natural Language Processing Opioid Misuse Classifier: Detection and Mitigation of Electronic Health Record Data Disadvantages across Racial Subgroups. *Journal of the American Medical Informatics Association* 28, 11 (Nov. 2021), 2393–2403. doi:10.1093/jamia/ocab148

[314] H. Tian, B. Liu, T. Zhu, W. Zhou, and P.S. Yu. 2024. MultiFair: Model Fairness With Multiple Sensitive Attributes. *IEEE Transactions on Neural Networks and Learning Systems* 36, 3 (2024), 1–14. doi:10.1109/TNNLS.2024.3384181

[315] Songül Tolan, Marius Miron, Emilia Gómez, and Carlos Castillo. 2019. Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law (Icail '19)*. Association for Computing Machinery, New York, NY, USA, 83–92. doi:10.1145/3322640.3326705

[316] M. Topaz, A. Davoudi, L. Evans, S. Sridharan, J. Song, S. Chae, Y. Barrón, M. Hobensack, D. Scharp, K. Cato, S.C. Rossetti, P. Kapela, Z. Xu, P. Gupta, Z. Zhang, M.V. Mcdonald, and K.H. Bowles. 2025. Building a Time-Series Model to Predict Hospitalization Risks in Home Health Care: Insights Into Development, Accuracy, and Fairness. *Journal of the American Medical Directors Association* 26, 2 (2025), 1–9. doi:10.1016/j.jamda.2024.105417

[317] A. Trang, K. Putman, D. Savani, D. Chatterjee, J. Zhao, P. Kamel, J.J. Jeudy, V.S. Parekh, and P.H. Yi. 2024. Sociodemographic Biases in a Commercial AI Model for Intracranial Hemorrhage Detection. *Emergency Radiology* 31, 5 (2024), 713–723. doi:10.1007/s10140-024-02270-w

[318] C. Trentz, J. Engelbart, J. Semprini, A. Kahl, E. Anyimadu, J. Buatti, T. Casavant, M. Charlton, and G. Canahuate. 2024. Evaluating Machine Learning Model Bias and Racial Disparities in Non-Small Cell Lung Cancer Using SEER Registry Data. *Health Care Management Science* 27, 4 (2024), 631–649. doi:10.1007/s10729-024-09691-6

[319] Anne M Tumlin, Diego Manzanas Lopez, Preston Robinette, Yuying Zhao, Tyler Derr, and Taylor T Johnson. 2024. FairNNV: The Neural Network Verification Tool for Certifying Fairness. In *Proceedings of the 5th ACM International Conference on AI in Finance (Icaif '24)*. Association for Computing Machinery, New York, NY, USA, 36–44. doi:10.1145/3677052.3698677

[320] J. Ugirumurera, E.A. Bensen, J. Severino, and J. Sanyal. 2024. Addressing Bias in Bagging and Boosting Regression Models. *Scientific Reports* 14, 1 (2024), 1–12. doi:10.1038/s41598-024-68907-5

[321] Ana Valdivia, Javier Sánchez-Monedero, and Jorge Casillas. 2021. How Fair Can We Go in Machine Learning? Assessing the Boundaries of Accuracy and Fairness. *International Journal of Intelligent Systems* 36, 4 (April 2021), 1619–1643. doi:10.1002/int.22354

[322] K. Van Busum and S. Fang. 2024. Bias Analysis of AI Models for Undergraduate Student Admissions. *Neural Computing and Applications* 37 (2024), 7785–7795. doi:10.1007/s00521-024-10762-6

[323] Michael Varley and Vaishak Belle. 2021. Fairness in Machine Learning with Tractable Models. *Knowledge-Based Systems* 215 (March 2021), N.PAG–N.PAG. doi:10.1016/j.knosys.2020.106715

[324] Sriram Vasudevan and Krishnaram Kenthapadi. 2020. LiFT: A Scalable Framework for Measuring Fairness in ML Applications. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (Cikm '20)*. Association for Computing Machinery, New York, NY, USA, 2773–2780. doi:10.1145/3340531.3412705

[325] M. Vega-Gonzalo and P. Christidis. 2022. Fair Models for Impartial Policies: Controlling Algorithmic Bias in Transport Behavioural Modelling. *Sustainability (Switzerland)* 14, 14 (2022), 1–23. doi:10.3390/su14148416

[326] M. Verger, C. Fan, S. Lallé, F. Bouchet, and V. Luengo. 2024. A Comprehensive Study on Evaluating and Mitigating Algorithmic Unfairness with the MADD Metric. *Journal of Educational Data Mining* 16, 1 (2024), 365–409. doi:10.5281/zenodo.12180668

[327] A. Vetrò, M. Torchiano, and M. Mecati. 2021. A Data Quality Approach to the Identification of Discrimination Risk in Automated Decision Making Systems. *Government Information Quarterly* 38, 4 (2021), 1–17. doi:10.1016/j.giq.2021.101619

[328] M. Vitek, A. Das, D.R. Lucio, L.A. Zanlorensi, D. Menotti, J.N. Khiarak, M.A. Shahpar, M. Asgari-Chenaghlu, F. Jaryani, J.E. Tapia, A. Valenzuela, C. Wang, Y. Wang, Z. He, Z. Sun, F. Boutros, N. Damer, J.H. Grebe, A. Kuijper, K. Raja, G. Gupta, G. Zampoukis, L. Tsochatzidis, I. Pratikakis, S.V. Aruna Kumar, B.S. Harish, U. Pal, P. Peer, and V. Struc. 2023. Exploring Bias in Sclera Segmentation Models: A Group Evaluation Approach. *IEEE*

Transactions on Information Forensics and Security 18 (2023), 190–205. doi:10.1109/TIFS.2022.3216468

[329] Clarice Wang, Kathryn Wang, Andrew Y. Bian, Rashidul Islam, Kamrun Naher Keya, James Foulds, and Shimei Pan. 2023. When Biased Humans Meet Debiased AI: A Case Study in College Major Recommendation. *ACM Transactions on Interactive Intelligent Systems* 13, 3, Article 17 (Sept. 2023), 29 pages. doi:10.1145/3611313

[330] H.E. Wang, J.P. Weiner, S. Saria, and H. Kharrazi. 2024. Evaluating Algorithmic Bias in 30-Day Hospital Readmission Models: Retrospective Analysis. *Journal of Medical Internet Research* 26, 1 (2024), 1–18. doi:10.2196/47125

[331] Haonan Wang, Ziwei Wu, and Jingrui He. 2024. FairIF: Boosting Fairness in Deep Learning via Influence Functions with Validation Set Sensitive Attributes. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (Wsdm '24)*. Association for Computing Machinery, New York, NY, USA, 721–730. doi:10.1145/3616855.3635844

[332] H. Echo Wang, Jonathan P. Weiner, Suchi Saria, Harold Lehmann, and Hadi Kharrazi. 2024. Assessing Racial Bias in Healthcare Predictive Models: Practical Lessons from an Empirical Evaluation of 30-Day Hospital Readmission Models. *Journal of Biomedical Informatics* 156 (Aug. 2024), N.PAG–N.PAG. doi:10.1016/j.jbi.2024.104683

[333] M. Wang, W. Deng, J. Hu, and S. Su. 2025. Marginal Debiased Network for Fair Visual Recognition. *Pattern Recognition* 158 (2025), 1–11. doi:10.1016/j.patcog.2024.111046

[334] R. Wang, P. Chaudhari, and C. Davatzikos. 2023. Bias in Machine Learning Models Can Be Significantly Mitigated by Careful Training: Evidence from Neuroimaging Studies. *Proceedings of the National Academy of Sciences of the United States of America* 120, 6 (2023), 1–3. doi:10.1073/pnas.2211613120

[335] R. Wang, P.-C. Kuo, L.-C. Chen, K.P. Seastedt, J.W. Gichoya, and L.A. Celi. 2024. Drop the Shortcuts: Image Augmentation Improves Fairness and Decreases AI Detection of Race and Other Demographics from Medical Images. *eBioMedicine* 102 (2024). doi:10.1016/j.ebiom.2024.105047

[336] T. Wang, K. Zhang, J. Cai, Y. Gong, K.-K.R. Choo, and Y. Guo. 2024. Analyzing the Impact of Personalization on Fairness in Federated Learning for Healthcare. *Journal of Healthcare Informatics Research* 8, 2 (2024), 181–205. doi:10.1007/s41666-024-00164-7

[337] X. Wang, C.-H. Chang, and C.C. Yang. 2024. Achieving Equity via Transfer Learning With Fairness Optimization. *IEEE Access* 12 (2024), 195229–195241. doi:10.1109/ACCESS.2024.3519465

[338] Y. Wang and L. Singh. 2021. Analyzing the Impact of Missing Values and Selection Bias on Fairness. *International Journal of Data Science and Analytics* 12, 2 (2021), 101–119. doi:10.1007/s41060-021-00259-z

[339] Yifei Wang, Liqin Wang, Zhengyang Zhou, John Laurentiev, Joshua R. Lakin, Li Zhou, and Pengyu Hong. 2024. Assessing Fairness in Machine Learning Models: A Study of Racial Bias Using Matched Counterparts in Mortality Prediction for Patients with Chronic Diseases. *Journal of Biomedical Informatics* 156 (Aug. 2024), N.PAG–N.PAG. doi:10.1016/j.jbi.2024.104677

[340] Yan Wang, Ruochi Zhang, Qian Yang, Qiong Zhou, Shengde Zhang, Yusi Fan, Lan Huang, Kewei Li, and Fengfeng Zhou. 2024. FairCare: Adversarial Training of a Heterogeneous Graph Neural Network with Attention Mechanism to Learn Fair Representations of Electronic Health Records. *Information Processing & Management* 61, 3 (May 2024), N.PAG–N.PAG. doi:10.1016/j.ipm.2024.103682

[341] Zichong Wang, Nripsuta Saxena, Tongjia Yu, Sneha Karki, Tyler Zetty, Israat Haque, Shan Zhou, Dukka Kc, Ian Stockwell, Xuyu Wang, Albert Bifet, and Wenbin Zhang. 2023. Preventing Discriminatory Decision-Making in Evolving Data Streams. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 149–159. doi:10.1145/3593013.3593984

[342] Solvejg Wastvedt, Joshua Snoke, Denis Agniel, Julie Lai, Marc N Elliott, and Steven C Martino. 2024. De-Biasing the Bias: Methods for Improving Disparity Assessments with Noisy Group Measurements. *Biometrics* 80, 4 (Dec. 2024), 1–11. doi:10.1093/biomtc/ujae155

[343] Z. Wei, Y. Zhou, Z. Li, M. Kulkarni, and Y. Zhang. 2024. Supporting Equitable and Responsible Highway Safety Improvement Funding Allocation Strategies – Why AI Prediction Biases Matter. *Accident Analysis and Prevention* 202 (2024), 10–15. doi:10.1016/j.aap.2024.107585

[344] E.F. Wong, A.K. Saini, E.E. Accortt, M.S. Wong, J.H. Moore, and T.J. Bright. 2024. Evaluating Bias-Mitigated Predictive Models of Perinatal Mood and Anxiety Disorders. *JAMA network open* 7, 12 (2024), e2438152. doi:10.1001/jamanetworkopen.2024.38152

[345] T. Wongvorachan, O. Bulut, J.X. Liu, and E. Mazzullo. 2024. A Comparison of Bias Mitigation Techniques for Educational Classification Tasks Using Supervised Machine Learning. *Information (Switzerland)* 15, 6 (2024), 1–25. doi:10.3390/info15060326

[346] Jiahui Wu, Saad Mohammad Abrar, Naman Awasthi, and Vanessa Frías-Martínez. 2023. Auditing the Fairness of Place-Based Crime Prediction Models Implemented with Deep Learning Approaches. *Computers, Environment & Urban Systems* 102 (June 2023), N.PAG–N.PAG. doi:10.1016/j.compenvurbsys.2023.101967

[347] Jiahui Wu, Enrique Frias-Martinez, and Vanessa Frias-Martinez. 2020. Addressing Under-Reporting to Enhance Fairness and Accuracy in Mobility-Based Crime Prediction. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems (Sigspatial '20)*. Association for Computing Machinery, New York, NY, USA, 325–336. doi:10.1145/3397536.3422205

[348] Ziwei Wu and Jingrui He. 2022. Fairness-Aware Model-Agnostic Positive and Unlabeled Learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1698–1708. doi:10.1145/3531146.3533225

[349] Sierra Wyllie, Ilia Shumailov, and Nicolas Papernot. 2024. Fairness Feedback Loops: Training on Synthetic Data Amplifies Bias. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2113–2147. doi:10.1145/3630106.3659029

[350] Jing Nathan Yan, Ziwei Gu, Hubert Lin, and Jeffrey M. Rzeszotarski. 2020. Silva: Interactively Assessing Machine Learning Fairness Using Causality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Chi '20)*. Association for Computing Machinery, New York, NY,

USA, 1–13. doi:10.1145/3313831.3376447

[351] M. Yan, M.J. Pencina, L.E. Boulware, and B.A. Goldstein. 2022. Observability and Its Impact on Differential Bias for Clinical Prediction Models. *Journal of the American Medical Informatics Association* 29, 5 (2022), 937–943. doi:10.1093/jamia/ocac019

[352] Shen Yan, Hsien-te Kao, and Emilio Ferrara. 2020. Fair Class Balancing: Enhancing Model Fairness without Observing Sensitive Attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (Cikm '20)*. Association for Computing Machinery, New York, NY, USA, 1715–1724. doi:10.1145/3340531.3411980

[353] J. Yang, A.A.S. Soltan, D.W. Eyre, and D.A. Clifton. 2023. Algorithmic Fairness and Bias Mitigation for Clinical Machine Learning with Deep Reinforcement Learning. *Nature Machine Intelligence* 5, 8 (2023), 884–894. doi:10.1038/s42256-023-00697-3

[354] J. Yang, A.A.S. Soltan, D.W. Eyre, Y. Yang, and D.A. Clifton. 2023. An Adversarial Training Framework for Mitigating Algorithmic Biases in Clinical Machine Learning. *npj Digital Medicine* 6, 1 (2023), 1–10. doi:10.1038/s41746-023-00805-y

[355] Ming Ying Yang, Gloria Hyunjung Kwak, Tom Pollard, Leo Anthony Celi, and Marzyeh Ghassemi. 2023. Evaluating the Impact of Social Determinants on Health Prediction in the Intensive Care Unit. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (Aies '23)*. Association for Computing Machinery, New York, NY, USA, 333–350. doi:10.1145/3600211.3604719

[356] P. Yang, I.A. Gregory, C. Robichaux, A.L. Holder, G.S. Martin, A.M. Esper, R. Kamaleswaran, J.W. Gichoya, and S.V. Bhavani. 2024. Racial Differences in Accuracy of Predictive Models for High-Flow Nasal Cannula Failure in COVID-19. *Critical Care Explorations* 6, 3 (2024), 1–10. doi:10.1097/CCE.0000000000001059

[357] Min-Hsuan Yeh, Blossom Metevier, Austin Hoag, and Philip Thomas. 2024. Analyzing the Relationship between Difference and Ratio-Based Fairness Metrics. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 518–528. doi:10.1145/3630106.3658922

[358] Sofia Yfantidou, Dimitris Spathis, Marios Constantinides, Athena Vakali, Daniele Quercia, and Fahim Kawsar. 2024. Using Self-Supervised Learning Can Improve Model Fairness. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Kdd '24)*. Association for Computing Machinery, New York, NY, USA, 3942–3953. doi:10.1145/3637528.3671991

[359] V. Yogarajan, G. Dobbie, S. Leitch, T.T. Keegan, J. Bensemann, M. Witbrock, V. Asrani, and D. Reith. 2022. Data and Model Bias in Artificial Intelligence for Healthcare Applications in New Zealand. *Frontiers in Computer Science* 4 (2022), 1–21. doi:10.3389/fcomp.2022.1070493

[360] C. Yuan, K.A. Linn, and R.A. Hubbard. 2023. Algorithmic Fairness of Machine Learning Models for Alzheimer Disease Progression. *JAMA Network Open* 6, 11 (2023), 1–14. doi:10.1001/jamanetworkopen.2023.42203

[361] Andres Felipe Zambrano, Jiayi Zhang, and Ryan S. Baker. 2024. Investigating Algorithmic Bias on Bayesian Knowledge Tracing and Carelessness Detectors. In *Proceedings of the 14th Learning Analytics and Knowledge Conference (Lak '24)*. Association for Computing Machinery, New York, NY, USA, 349–359. doi:10.1145/3636555.3636890

[362] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (Aies '18)*. Association for Computing Machinery, New York, NY, USA, 335–340. doi:10.1145/3278721.3278779

[363] Jiang Zhang, Ivan Beschastnikh, Sergey Mechtaev, and Abhik Roychoudhury. 2022. Fair Decision Making via Automated Repair of Decision Trees. In *Proceedings of the 2nd International Workshop on Equitable Data and Technology (FairWare '22)*. Association for Computing Machinery, New York, NY, USA, 9–16. doi:10.1145/3524491.3527306

[364] Jindi Zhang, Luning Wang, Dan Su, Yongxiang Huang, Caleb Chen Cao, and Lei Chen. 2023. Model Debiasing via Gradient-Based Explanation on Representation. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (Aies '23)*. Association for Computing Machinery, New York, NY, USA, 193–204. doi:10.1145/3600211.3604668

[365] Linying Zhang, Lauren R. Richter, Yixin Wang, Anna Ostropolets, Noémie Elhadad, David M. Blei, and George Hripcsak. 2024. Causal Fairness Assessment of Treatment Allocation with Electronic Health Records. *Journal of Biomedical Informatics* 155 (July 2024), N.PAG–N.PAG. doi:10.1016/j.jbi.2024.104656

[366] W. Zhang, F. He, and S. Zhang. 2023. A Novel Fairness-Aware Ensemble Model Based on Hybrid Sampling and Modified Two-Layer Stacking for Fair Classification. *International Journal of Machine Learning and Cybernetics* 14, 11 (2023), 3883–3896. doi:10.1007/s13042-023-01870-1

[367] Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. 2022. Towards Fair Classifiers without Sensitive Attributes: Exploring Biases in Related Features. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (Wsdm '22)*. Association for Computing Machinery, New York, NY, USA, 1433–1442. doi:10.1145/3488560.3498493

[368] W. Zheng, L. Lin, X. Wu, and X. Chen. 2024. An Empirical Study on Correlations between Deep Neural Network Fairness and Neuron Coverage Criteria. *IEEE Transactions on Software Engineering* 50, 3 (2024), 391–412. doi:10.1109/TSE.2023.3349001

[369] Yunhan Zheng, Shenhao Wang, and Jinhua Zhao. 2021. Equality of Opportunity in Travel Behavior Prediction with Deep Neural Networks and Discrete Choice Models. *Transportation Research Part C: Emerging Technologies* 132 (Nov. 2021), N.PAG–N.PAG. doi:10.1016/j.trc.2021.103410

[370] Miri Zilka, Riccardo Fogliato, Jiri Hron, Bradley Butcher, Carolyn Ashurst, and Adrian Weller. 2023. The Progression of Disparities within the Criminal Justice System: Differential Enforcement and Risk Assessment Instruments. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1553–1569. doi:10.1145/3593013.3594099

Received 27 January 2026