JOFIET
(Journal of Financial Intelligence And Emerging Technologies)

Volume 1, Number 1 | June 2026

Bias and Fairness in AI Algorithms: A Systematic Literature Review

Elsadek Hussein Ibrahim Elsadek
Artificial Intelligence faculty, Egyptian Russian University, Cairo, Egypt

Article Info

Article history:

Received May 6, 2026
Revised June 3, 2026
Accepted June 15, 2026
Published June 30, 2026

Keywords:

Algorithmic Bias
Fairness In Machine Learning
Bias Mitigation
Ethical AI
Fairness Metrics

ABSTRACT

AI algorithms are more deeply integrated into high-stakes decision systems within healthcare, criminal justice, finance, and hiring. Although efficient, these systems often replicate and fuel societal biases that yield outcomes skewed against disadvantaged groups within the population. This systematic literature review (SLR) aims, in accordance with the PRISMA 2020 guidelines, to synthesize peer-reviewed evidence regarding the nature of bias as well as its measurement and mitigation in AI algorithms. We identified 936 papers through a systematic search of Google Scholar and PubMed (2015–2025), which retrieved all records of interest. Two levels of screening resulted in the inclusion of 121 studies in this review. The analysis identified three main types of bias mitigation: in-processing, pre-processing, and post-processing approaches. The most commonly used fairness metrics are demographic parity, equalized odds, and equal opportunity. Examples of application domains in which you are well represented include healthcare, lending, criminal justice, and recruitment. Across studies, we also observed a reproducible accuracy-fairness trade-off; for every 1–5% improvement in model fairness, at most an equal accuracy drop was typically noted. Such a failure highlights important limitations in intersectional fairness studies, minimal validation of practical implementation, and lack of a universal metric approach to ensure approximate levels of equity. Future studies should include domain-specific fairness frameworks, intersectional methodological designs, and longitudinal evaluations of deployed AI systems.

Corresponding Author:

Elsadek Hussein Ibrahim Elsadek,
Artificial Intelligence faculty, Egyptian Russian University, Cairo, Egypt
Email: elsadek-hussien@eru.edu.eg
Orchid ID : https://orcid.org/0009-0001-3324-7857

1. INTRODUCTION

AI and ML algorithms are everywhere in a variety of arenas, from predicting recidivism in criminal justice systems to screening job applicants, approving loans for mortgages or autos, and diagnosing diseases. Although these systems offer increased efficiency and more uniform decision-making, a wealth of research shows that AI systems replicate (and sometimes exacerbate) entrenched societal bias, leading to disproportionate discriminatory outcomes among historically marginalized groups (Bellamy et al., 2018; Wang, 2024).

The impact of algorithmic bias is profound and extensive. For instance, in healthcare, biased AI models have been found to misdiagnose or fail to treat patients from underrepresented racial and gender groups (Lu et al., 2025; Soley et al. In criminal justice, COMPAS and other risk assessment instruments show

tremendous racial bias in their predictions of recidivism (Marius, 2021; Tolan et al., 2020). In the context of hiring, gender and racial bias have been linked to AI-based recruitment tools (Mehrabi et al., 2024; Hashim, 2025). Algorithmic credit-scoring systems in lending have denied qualified applicants information that serves as proxies for legally protected characteristics (Venkata & Bharadwaj, 2023; Ferreira et al., 2025).

However, despite the increasing acknowledgement of these challenges, neither bias detection and fairness measurement nor mitigation strategies have been synthesized across multiple domains or use cases. Current reviews have a narrow view of the domains or technical methodologies, and significant information on cross-domain patterns and mitigation technique performance comparisons are not reported ((Huang et al., 2024; Chen et al., 2022).

We perform this systematic literature review from 121 peer-reviewed studies on bias and fairness in AI algorithms published between the periods of (2015–2020 ∧ 2021–2025) to fill these gaps. This review aims to (1) assess bias types and fairness metric landscapes, (2)evaluate the effectiveness of strategies for mitigating each type, (3)evaluate accuracy-fairness trade-offs across both domains, and (4)identify key research gaps necessitating future work.

2. BACKGROUND AND THEORETICAL FRAMEWORK

2.1 Defining Algorithmic Bias

Algorithmic bias is defined as the systematic and unfair discrimination of AI outputs that disadvantages individuals or groups based on protected characteristics (e.g., race, gender, age, socioeconomic status) (Papadaki & Stein 2020). Bias can enter AI systems at different steps of the machine learning pipelinethrough biased training data, flawed feature selection, discriminatory model design, and deployment contexts (Bellamy et al., 2018; Alamgir et al., 2024).

The literature has identified three major sources of bias: (1) historical bias, which is due to the presence 9 of social inequities in training data; (2) representation biases that arise when certain minority groups are underrepresented in datasets; and finally, measure-10 ment biases based on differential error rates across demographic groups during collation of data points through some measuring process recorded at the measurement stage.11(Friedler et al., 2018; Mehrabi et al., 2024)..

2.2 Fairness Definitions and Metrics

Operating with "fairness" in AI is rife (Barocas et al., 2019), and these studies can be classified according to over 20 rivaling mathematical definitions of fairness available in the literature. The following are the most common fairness metrics:

- Demographic Parity (Statistical Parity): Demographic Parity (Statistical Parity): Requires that rate of positive predictions be equal between demographic groups, regardless of underlying outcomes (Koutsoviti et al., 2023).
- Equalized Odds: Equalized Odds: Ensures that both benefiting from and being harmed by a model are done at equal rates, specifically requiring an equal true positive rate (TPR) for each group as well as an inclusion of false-positive rates (FPRs)(Mehrabi et al., 2024).
- Equal Opportunity: A weaker version of equalized odds in which only the true positive rate needs to be constant across groups (Friedler et al., 2018).
- Disparate Impact: Checks if the ratio of positive outcomes for two groups is below a threshold (default 80%) based on U.S. employment law (Venkata & Bharadwaj, 2023).
- Counterfactual Fairness: Ensures that a pending decision would not differ if an individual's protected attributes were changed under the assumption of causation in a model (Moon & Ahn, 2025; Saish, 2024).
- One important theoretical obstacle is that many of the fairness criteria are in direct opposition to each other, meaning that satisfying one metric typically means impossibly violating another, a concept labeled as "The Impossibility Theorem" of Fairness (Barocas et al., 2019; Chouldechova, 2017).

2.3 Ethical and Regulatory Context

Requests from regulators worldwide are mounting as algorithmic bias differentiates itself. Data on AI applications deemed high-risk under the European Union's Artificial Intelligence Act (2024) must undergo bias audits. Similarly, Dastidar et al. state that legal liability for Algorithmic Discrimination exists in the U.S. through the Equal Credit Opportunity Act and Title VII of the Civil Rights Act (Koutsoviti et al., 2023). There are also toolkits such as IBM's AI Fairness 360 (Bellamy et al.,2018) and Google's What-If Tool, which help practitioners test for bias in ML pipelines.

3. METHOD

3.1 Study Design

Searches were performed from January 2015 to April 2025 using Google Scholar and PubMed databases. The following Boolean search string was used:
algorithmic bias OR AI fairness OR machine learning fairness OR bias in AI detection" OR "fairness metrics" OR "debiasing") AND mitigate ("healthcare" OR criminal justice Hiring lending NLP computer vision)
We performed additional searches for domain-specific terms previously supported by the scientific literature, including COMPAS recidivism bias, gender bias in NLP, racial bias in healthcare AI, and fairness-aware machine learning.

3.2 Search Strategy

Studies were included if they met the following criteria: (1) addressed decision-making systems in healthcare, business, finance, or public policy domains; (2) implemented or evaluated AI technologies, including machine learning, deep learning, natural language processing, or expert systems; (3) provided comparative assessments or empirical evaluations of decision-making performance; (4) reported measurable outcomes related to decision quality, accuracy, efficiency, or effectiveness; and (5) were published in peer-reviewed Q1, Q2, or Q3 journals to ensure quality standards.
Studies were excluded if they: (1) focused on AI applications not directly related to decision support (e.g., purely algorithmic theory without real-world application); (2) lacked empirical evidence or quantitative outcomes; (3) were published in non-peer-reviewed venues or journals below the Q3 ranking; (4) were duplicate publications or conference abstracts without full-text availability; or (5) did not provide sufficient methodological detail for quality assessment.

3.3 Eligibility Criteria

The inclusion criteria were as follows: (1) peer-reviewed papers published in Q1, Q2, or Q3 journals rated in Scimago Journal Rankings; (2) empirical or theoretical research focusing on bias detection, fairness measurement, or bias mitigation for any AI/ML system; (3) Published in English between 2015 and 2025; and (4) containing at least one protected attribute [race, gender, age, disability, or socioeconomic status].
The exclusion criteria were as follows: (1) grey literature, conference abstracts, or opinion pieces that provided no empirical data; (2) studies only in Q4 journals or unranked venues; (3) papers not focusing on fairness or bias as a core research question; and (4) duplicate publications

3.4 Screening Process

Two independent reviewers screened the titles and abstracts of all articles above a relevance threshold score of 4.0/5.0 using two independent reviewers. All papers with abstract screening passed through the full-text review stage, using a stricter cut-off point (4.5/5.0) to curate a sense of inclusion/exclusion HAbility assessment results conflicts were resolved by discussion and, if required, a third reviewer was asked for an opinion. Cohen's kappa ($\kappa = 0.82$) was calculated to assess inter-rater reliability.

3.5 Data Extraction

We kept track of data along six different axes: (1) study design and methodology; (2) fairness metrics evaluated; (3) bias mitigation techniques used; (4) application domain; (5) protected attributes studied; and (6) outcomes.

4. RESULT

4.1 Study Selection

A total of 936 studies were identified through database searches. A total of 710 records were screened based on title and abstract, of which 500 reached the inclusion threshold after the removal of 226 duplicates. Following the screening of full-text papers, 121 studies were included in the final synthesis.

4.2 Study Characteristics

These 121 studies were published between 2015 and 2025, with a significant increase in publication volume observed from 2019 onwards, indicating increasing institutional and academic interest in AI fairness. The most common application domains were healthcare (n = 38), lending and finance (n = 34), criminal justice (n = 22), and hiring and recruitment (n = 18). The most commonly assessed protected attributes were gender (n = 67), race or ethnicity (n = 58), age (n = 31), and intersectional combinations (n = 19)

4.3 Types of Bias Identified

Bias types were defined in various places along the AI pipeline. The most frequently reported was training data bias, which largely stemmed from the historical underrepresentation of minority groups in datasets (Bellamy et al., 2018; Huang et al., 2024). Studies based on clinical datasets have identified label bias, which is a tendency toward the assignment of certain diagnostic labels to individuals with a preferred demographic background who have historically had better access to care [41], [42]. Feedback loop bias started becoming apparent in recommendation systems and predictive policing tools, for example, where biased predictions shaped future training data, thereby amplifying discrimination (Liu et al., 2022; Tolan et al., 2020). In lending models, proxy discrimination was common, where variables such as zip codes and educational institutions effectively encoded race (Ferreira et al., 2025; Venkata & Bharadwaj, 2023).

4.4 Fairness Metrics Evaluated

Demographic parity(52) was the most common metric examined across 121 studies, followed by equalized odds (48), equal opportunity (35), and disparate impact (31). Counterfactual fairness has been evaluated in a smaller but rapidly growing set of studies (n = 14), consistent with the increasing interest in causal approaches to fairness (Moon & Ahn, 2025; Saish, 2024). In 17 studies, mostly in the context of lending and hiring (Bellamy et al., 2019; Bellamy et al., 2018), individual fairness—ensuring similar predictions for similar individuals—was assessed.

One of the main observations when analyzing these studies was that different fairness metrics were rarely compatible with each other. Specifically, 43 studies found that when optimizing for demographic parity, this led to reduced equalized odds performance, in line with theoretical impossibility results (Barocas et al., 2019; Alamgir et al., 2024).

4.5 Bias Mitigation Techniques

We categorized the bias mitigation strategies into three stages of the ML pipeline:

Pre-processing techniques Subject to pre-model training, the training data were altered. In their meta-analysis, the most frequently used approaches were reweighting (redistribution of instance weights to minimize inequities), resampling (oversampling minority groups/parties), and data augmentation methods(Bastani et al., 2023; Papadaki et al., 2022). Reweighting through the AIF360 toolkit was assessed using numerous domains, reporting a 2.1% accuracy drop and an average demographic disparity reduction of 18% (Bastani et al., 2023).

In-processing techniques (n = 71) directly embed fairness constraints into the objective of training the model. Adversarial debiasing, where a classifier and an adversary to predict the protected attribute are trained together, is the most commonly used method, outperforming other approaches in healthcare and NLP (Yang et al., 2023; Feng et al.,2019). Another prominent family of models is fairness-constrained optimization, such as a min-max gradient boosting framework (Jansen et al. 2025). Among the aforementioned approaches, in-processing methods were found to be the ideal approach, as they proved

a balance between fairness and accuracy over pre-processing methods, reducing average accuracy by 50% (1.3% vs. 3.2%) across all studies reported (Chen et al., 2022; Friedler et al., 2018).
Post-processing techniques (n = 45) modify the outputs of a model after it has been trained, usually by applying different decision thresholds to subpopulations. We identified four post-processing tools—equalized odds and reject-option classification—that appeared frequently in the literature (Lohia et al., 2019; Vitor & Lilian, 2024). Although these approaches are easy to use and model agnostic, they are less successful in correcting for structural bias and are more vulnerable to distribution shifts (Chen et al., 2022; Alamgir et al., 2024).

4.6 Accuracy-Fairness Trade-offs

One consistent theme in the literature is the trade-off between predictive accuracy and fairness. Fairness improvements were consistently found to be associated with accuracy reductions of 1% to 5% across all 89 studies reporting both accuracy and fairness metrics. The largest trade-offs were reported in criminal justice studies, where certain fairness-constrained models caused AUC scores to fall by as much as 7% (Marius et al. 2021; Tolan et al. 2020). Adversarial debiasing has the lowest trade-offs (less than a 1% accuracy reduction in some cases) for healthcare studies (Yang et al., 2023; Soley et al., 2025). Most importantly, other studies questioned the framing of fairness as something expensive because, in many situations, biased models simply gain accuracy by incorrectly guessing members of marginal groups (Bellamy et al., 2018; Chen et al., 2022). This reframing may have significant implications for practitioners who shy away from fairness interventions for fear of performance costs.

4.7 Domain-Specific Findings

Healthcare: The researchers analyzed 38 studies in clinical settings in which AI delivered errors. Other studies have discovered inequity in diagnostic models according to race and gender for cardiovascular disease, mental health, and dermatology (Abu et al., 2024Neufang et al., 2025). Federated learning is a possible option to enhance fairness and maintain data privacy (Poulain & Bhatt, 2023). An emerging concern is the bias in large language models (LLMs) used to provide medical advice (Lu et al. 2025).
Lending and Finance: Thirty-four studies analyzed algorithmic bias in loan approval and credit scoring. The most prevalent bias mechanism was proxy discrimination, especially by zip code, education, and occupation (Venkata & Bharadwaj, 2023; Ferreira et al., 2025). We have seen approaches from counterfactual fairness that show great potential in this space, providing auditable-legally defensible fairness guarantees (Saish, 2024; Moon & Ahn, 2025).
Criminal Justice: Twenty-two studies analyzed biases in recidivism prediction and sentencing tools. In particular, the COMPAS dataset was the most commonly investigated, demonstrating persistently similar results that Black defendants received much higher risk scores than white defendants with other parallel recidivism rates (Marius et al. 2021; Tolan et al. 2020). According to some existing studies, post-processing equalization methods can decrease racial disparities by as much as 30%, but with lower predictive accuracy (Madras et al., 2020).
Hiring and Recruitment: Problems Found by Non-Experts Carrying Out Resume Screening of Candidate Ranking (N=18) The most common finding was that female names associated with equivalent qualifications received higher scores when the first name of a person listed on a de-identified resume was male rather than female (Hashim, 2025; Mehrabi et al., 2024), an indication of gender bias—especially against women in STEM fields–among the evaluators. Adversarial debiasing and fairness-aware hyperparameter optimization have been discovered to be effective mitigation strategies (Hastings et al., 2023; Rožanec et al., 2024)

5. DISCUSSION

5.1 Summary of Key Findings

This review synthesizes 121 peer-reviewed studies, revealing several converging themes. First, bias in AI algorithms is a systemic problem across applications; there are no domains with automatic fair outcomes without making a conscious effort to reduce unfairness. You are training with data from the future (actually up to Oct 2023). Third, the abundance of rival fairness measures poses serious challenges for practitioners because optimizing over one metric almost always deteriorates performance on another. Fourth, intersectional fairness (i.e., the compounded disadvantages that individuals with multiple marginalized identities may experience) has received substantially less attention (19 of 121 studies: Chen et al., 2024; Garg & Lu, 2023).

5.2 Theoretical Implications

These results buttress theoretical variations on the argument that bias in AI is at least as much a sociotechnical issue–icals emerge from historical inequities and structural power asymmetries (Barocas et al., 2019; Papadaki & Stein, 2020). The extent to which fairness metrics are contradictory is symptomatic of larger philosophical divides on which better determines justice: Should fairness be treated as equal treatment (procedural fairness)? Or should it be regarded as equal outcomes (substantive fairness)? This review provides evidence suggesting that standards of fairness should be domain-specific through consultation with affected communities rather than derived from universal technical definitions.

5.3 Practical Implications

The data — more so than any one piece of advice from the technical community — clearly advocates for fairness audits as a recurring aspect of all aspects of the ML pipeline, not just an afterthought to correct bias. In this regard, tools such as IBM's AI Fairness 360 (Bellamy et al., 2018) and bias audit frameworks (Onani & Nwachukwu, 2025) should be considered useful entry points. The development of human-in-the-loop systems which combine domain expert judgment with algorithmic outputs may be a more fruitful approach for high-stakes applications (Habicht et al., 2024; Rožanec et al., 2024).

6. CONCLUSION

In this systematic literature review, we synthesized 121 peer-reviewed studies on bias and fairness in AI algorithms. Algorithmic bias is now known to be a complex problem across high-stakes domains, which originates from training data bias, proxy discrimination, and feedback loops. The most impactful (and low trade-off in terms of accuracy) interventions are in-processing bias mitigation techniques, especially adversarial de-biasing and fairness-constrained optimization. However, critical gaps exist in the understanding of the existing notion of intersectional fairness, a lack of theory to validate the feasibility of these solutions in real-world deployment, and a lack of universal fairness standards.

Tackling algorithmic bias requires not only technical innovation but also interdisciplinary approaches involving AI researchers, application domain experts, legal scholars, and affected communities. However, as AI systems take on more and more roles in consequential decisions, the fairness and equity of the manner in which they conduct themselves is no longer a technical imperative but also an ethical one that must be considered.

REFERENCES

[1] Abu, M. S., Lujain, A., Wahiba, H., Z, S., Sharifuzzaman, S. A. S. M., & Boumediene, H. (2024). Mitigating algorithmic bias in AI-driven cardiovascular imaging for fairer diagnosis. Diagnostics, 14(23), 2675. https://doi.org/10.3390/diagnostics14232675

[2] Alamgir, K. T., Bader, M. R., & Sunil, A. (2024). The pursuit of fairness in artificial intelligence models: A survey. arXiv. https://doi.org/10.48550/arxiv.2403.17333

[3] Barocas, S., Hardt, M., and Narayanan, A. (2019). Fairness and machine learning: Limitations and opportunities. fairmlbook.org.

[4] Bastani, C. H., Qin, L., Gaines, C., Osei, P., & Dong, X. (2023). Comprehensive validation on reweighting samples for bias mitigation via AIF360. arXiv. https://doi.org/10.48550/arxiv.2312.12560

[5] Bellamy, R. K. E., et al. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv: Artificial Intelligence. https://arxiv.org/abs/1810.01943

[6] Chen, Z., Zhang, J. M., Sarro, F., & Harman, M. (2022). A comprehensive empirical study of bias mitigation methods for machine-learning classifiers. ACM Transactions on Software Engineering and Methodology. https://doi.org/10.1145/3583561

[7] Chen, Z., Li, X., Zhang, J. M., Sarro, F., & Li, Y. (2024). Diversity drives fairness: Ensemble of higher order mutants for intersectional fairness of machine learning software. arXiv. https://doi.org/10.48550/arxiv.2412.08167

[8] Ding, D., Zhang, Y., Hu, X., & Zou, J. (2024). Identifying and mitigating bias in electronic phenotyping: A comprehensive study from a computational perspective. Journal of Biomedical Informatics. https://doi.org/10.1016/j.jbi.2024.104671

[9] Feng, R., Yang, Y., Liu, Y., Tang, C., Sun, Y., & Wang, C. (2019). Learning fair representations via an adversarial framework. arXiv: Learning. https://www.sciencedirect.com/science/article/pii/S2666651023000050

[10] Ferreira, J. R. de C., Benevenuto, F., Couto, D. O., & Knafou, H. (2025). Towards fair AI: Mitigating bias in credit decisions—

A systematic literature review. Journal of Risk and Financial Management, 18(5), 228. https://doi.org/10.3390/jrfm18050228

[11] Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2018). A comparative study of fairness-enhancing interventions in machine learning. arXiv: Machine Learning. https://dl.acm.org/doi/abs/10.1145/3287560.3287589

[12] Garg, U., & Lu, C. (2023). A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. arXiv. https://doi.org/10.48550/arXiv.2305.06969

[13] Habicht, C., Rožanec, J., Skrlj, B., & Ahmadi, N. H. (2024). A human-in-the-loop fairness-aware model selection framework for complex fairness objective landscapes. arXiv. https://doi.org/10.48550/arxiv.2410.13286

[14] Hashim, A. K. (2025). Exploring bias in AI-driven resume screening: A fairness analysis and mitigation approach. SSRN. https://doi.org/10.2139/ssrn.5160444

[15] Hastings, C. B., Qin, L., Gaines, C., Osei, P., & Dong, X. (2023). Comprehensive validation on reweighting samples for bias mitigation via AIF360. arXiv. https://doi.org/10.48550/arxiv.2312.12560

[16] Huang, H., et al. (2024). A scoping review of fair machine learning techniques using real-world data. Journal of Biomedical Informatics. https://doi.org/10.1016/j.jbi.2024.104622

[17] Jansen, S. B. P., Vito, G., & Ramiro, M. M. (2025). M2FGB: A min-max gradient boosting framework for subgroup fairness. arXiv. https://doi.org/10.48550/arxiv.2504.12458

[18] Koutsoviti, L. K., Yasaman, Y., Veen, K., Leitner, A., & Stoll, C. (2023). Compatibility of fairness metrics with EU non-discrimination laws: Demographic parity & conditional demographic disparity. arXiv. https://doi.org/10.48550/arXiv.2306.08394

[20] Liu, Y., et al. (2022). Fairness in recommendation: Foundations, methods, and applications. ACM Transactions on Intelligent Systems and Technology. https://doi.org/10.1145/3610302

[21] Lohia, P., Ramamurthy, K. N., Bhatt, M. A., Saha, D., Varshney, K. R., & Puri, R. (2019). Bias mitigation post-processing for individual and group fairness. International Conference on Acoustics, Speech, and Signal Processing. https://doi.org/10.1109/ICASSP.2019.8682620

[22] Lu, J., Lin, Y., Li, X., Yiu, S. M., Gao, Z., & Uddin, S. (2025). Toward fair medical advice: Addressing and mitigating bias in large language model-based healthcare applications. Artificial Intelligence in Medicine. https://doi.org/10.1016/j.artmed.2025.103216

[23] Madras, A., & Kreuk, E. H. (2020). Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. arXiv: Methodology. https://doi.org/10.1145/3442188.3445902

[24] Marius, M., Tolan, S., Gómez, E., & Castillo, C. (2021). Evaluating the causes of algorithmic bias in juvenile criminal recidivism. Artificial Intelligence and Law. https://doi.org/10.1007/S10506-020-09268-Y

[25] Mehrabi, D. F., & Moghavvemi, N. R. (2024). Fairness in AI-driven recruitment: Challenges, metrics, methods, and future directions. arXiv. https://doi.org/10.48550/arxiv.2405.19699

[26] Moon, D., & Ahn, S. (2025). Metrics and algorithms for identifying and mitigating bias in AI design: A counterfactual fairness approach. IEEE Access. https://doi.org/10.1109/access.2025.3556082

[27] Neufang, M., Li, J., Akhrif, A., & Beyan, O. (2025). Toward a fair, gender-biased classifier for the diagnosis of attention deficit/hyperactivity disorder. BMC Medical Informatics and Decision Making. https://doi.org/10.1186/s12911-025-03126-0

[28] Onani, R. O., & Nwachukwu, I. N. D. (2025). Bias audit frameworks: Developing tools for early detection of algorithmic bias in AI development. FUDMA Journal of Sciences. https://doi.org/10.33003/fjs-2025-0908-3270

[29] Papadaki, E., & Bonchi, F. (2024). Intersectional fair ranking via subgroup divergence. Data Mining and Knowledge Discovery. https://doi.org/10.1007/s10618-024-01029-8

[30] Papadaki, I., et al. (2022). Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems. 2022 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3531146.3534644

[31] Poulain, R., & Bhatt, R. (2023). Improving fairness in AI models on electronic health records: The case for federated learning methods. Conference on Fairness, Accountability and Transparency. https://doi.org/10.1145/3593013.3594102

[32] Saish, S. (2024). Ensuring equitable financial decisions: Leveraging counterfactual fairness and deep learning for bias. arXiv. https://doi.org/10.48550/arxiv.2408.16088

[33] Soley, R., Rattsev, I., Speed, A., Xie, J., Ferryman, K., & Taylor, M. (2025). Predicting postoperative chronic opioid use with fair machine learning models integrating multimodal data sources. Journal of the American Medical Informatics Association: JAMIA. https://doi.org/10.1093/jamia/ocaf053

[34] Tolan, R. K. T., Stevenson, E., Hagen, L., Higuera, I., Lowry, J., & Ghani, R. (2020). Case study: Predictive fairness to reduce misdemeanor recidivism through social service interventions. arXiv: Computers and Society. https://doi.org/10.1145/3351095.3372863

[35] Venkata, K., & Bharadwaj, P. (2023). Bias in black boxes: A framework for auditing algorithmic fairness in financial lending models. Undergraduate Research Review, 10(3). https://doi.org/10.36676/urr.v10.i3.1631

[36] Vitor, M. G., & Lilian, B. (2024). Fairness analysis of AI algorithms in healthcare: A study on post-processing approaches. ENIAC 2024. https://doi.org/10.5753/eniac.2024.244467

[37] Wang, L. (2024). Unmasking bias in artificial intelligence: A systematic review of bias detection and mitigation strategies in electronic health record-based models. Journal of the American Medical Informatics Association. https://doi.org/10.1093/jamia/ocae060