Open Access Library Journal
2025, Volume 12, e13340
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Adversarial Debiasing for Bias Mitigation in Healthcare AI Systems: A Literature Review

Joshua Waithira[1], Ruth Chweya[1], Ratemo Makiya Cyprian[2]

[1]Department of Computing Sciences, Kisii University, Kisii, Kenya
[2]Department of Computing and Information Sciences, Maasai Mara University, Narok, Kenya
Email: Joshuawaithira@gmail.com

How to cite this paper: Waithira, J., Chweya, R. and Cyprian, R.M. (2025) Adversarial Debiasing for Bias Mitigation in Healthcare AI Systems: A Literature Review. *Open Access Library Journal*, **12**: e13340.
https://doi.org/10.4236/oalib.1113340

Received: March 25, 2025
Accepted: May 28, 2025
Published: May 31, 2025

Abstract

The application of artificial intelligence (AI) in healthcare has tremendous potential for improving diagnostic precision and optimizing treatment and patient care. However, increasing dependence on such tools brings up urgent questions regarding the amplification of existing biases, which may detract from their ability to improve fair clinical decision-making. Adversarial debiasing, a method that utilizes fairness measures by contrasting a core predictive model with an adversarial network to reduce the influence of sensitive features, has emerged as an effective way of mitigating bias in AI systems. This review combines findings from 25 studies on several areas, encompassing the technical elements of adversarial learning and its practical applications in healthcare. The review offers extensive data and thoroughly assesses technological, ethical, and practical issues. This study reveals that adversarial debiasing improves fairness indicators and presents significant trade-offs, including reduced sensitivity and interpretability. We conclude with recommendations for future research avenues, encompassing prospective multicenter trials, adaptive training methodologies, hybrid debiasing strategies, and formulating standardized regulatory frameworks.

Subject Areas

Artificial Intelligence

Keywords

Adversarial Debiasing, Healthcare AI, Diagnostic Imaging, Bias Mitigation, Fairness

1. Introduction

Artificial intelligence is now a vital component of healthcare in today's day and

age, with immense advancements in diagnostic capabilities, predictive modeling, and decision support [1] [2]. However, as AI systems are used to carry out such tasks as reading medical images, predicting patient outcomes, or making treatment recommendations, overwhelming evidence now exists that such systems have the potential to inadvertently perpetuate or even amplify existing biases. Bias in healthcare algorithms can stem from unbalanced training data, historically biased datasets, or defects in algorithm design, resulting in inequities in patient diagnosis and care. For example, AI models for dermatology have been shown to underperform in detecting melanoma on darker skin tones due to training on predominantly lighter-skinned images [3]. Similarly, cardiovascular risk prediction tools have underestimated heart disease risk in Black patients, leading to inequities in preventive care [4]. Such issues are particularly concerning in diagnostic imaging, where AI analysis aids radiologists in interpreting complex imaging techniques such as X-rays, CT scans, MRIs, and PET scans. Even slight biases in these applications can lead to serious outcomes, including misdiagnosis or disparate treatment among different population subgroups.

This study aims to investigate the effectiveness of adversarial debiasing mechanisms in diagnostic imaging. It seeks to evaluate how well these mechanisms mitigate bias. The emerging solution of adversarial debiasing integrates fairness constraints within training procedures to tackle these challenges. The training of the main model is accompanied by an adversarial network that detects insights about sensitive attributes, including race, socioeconomic status, and gender. The model receives penalties when it successfully recovers the protected attributes in its outputs. Through bias-by-design strategies, researchers aim to establish hidden features that do not contain protected attributes while keeping outputs clinically relevant [5].

This literature review performs a critical assessment of the efficacy of adversarial debiasing in healthcare AI. Although initial work indicates that adversarial techniques can enhance fairness metrics, the majority observe trade-offs with vital clinical performance measures, such as diagnostic accuracy [6]. Moreover, although diagnostic imaging research has expanded exponentially in recent years, the majority have used retrospective datasets that may not reflect clinical diversity. Additionally, the relationship between healthcare applications and diagnostic imaging in particular requires careful clarification. This review differentiates findings pertinent to overarching clinical predictive analytics from those specifically affecting imaging diagnostics.

This review compiled and synthesized around 25 studies sourced from the Dimensions and Scopus databases (covering 2016-2025) that focus on adversarial debiasing. It does this by explaining the ideas behind these methods, giving a detailed summary of the research results, discussing the methodological limitations, and looking at the bigger effects on healthcare policy and clinical practice. The main objective is to create a narrative that is both complete and easy to understand. This will help guide future research projects and promote fair and ethical AI development.

2. Methods and Materials

This systematic review adhered to the PRISMA 2020 guidelines [7] to ensure transparency and scientific rigor. The review process involved identifying, evaluating, and synthesizing studies focused on adversarial debiasing in healthcare AI, particularly in diagnostic imaging.

2.1. Search Strategy

A comprehensive search was conducted using the Dimensions and Scopus databases. Peer-reviewed articles published in English between 2016 and 2025 were included.

The search strategy utilized relevant keywords and Boolean operators to maximize coverage. Search terms included: "Adversarial debiasing", "AI fairness", "Diagnostic imaging", "Healthcare bias", and "Healthcare AI". Manual and automated searches were undertaken to ensure robustness. Sources included journals, conference proceedings, and unpublished studies related to healthcare AI. The detailed strategy ensured the replicability of the review.

2.2. Study Selection Process

The PRISMA flowchart, shown in **Figure 1**, illustrates the study selection process. An initial search yielded 150 articles. Titles and abstracts were screened independently by two reviewers based on the predefined criteria. Studies that failed to meet inclusion criteria were excluded (See **Table 1**). Following a detailed full-text review, 25 studies were selected for inclusion. This systematic selection ensured a balance of quality and relevance, minimizing bias.

Table 1. Inclusion and exclusion measure.

Inclusion Criteria	Exclusion Criteria
Peer-reviewed articles published in English between 2016 and 2025.	Studies not related to adversarial debiasing or healthcare.
Directly or indirectly answers the research Question.	Lacks the relationship to the defined research inquiries of the study.
Research explicitly addresses adversarial debiasing in clinical or diagnostic environments.	Non-peer-reviewed publications (e.g., editorials, opinion pieces).
Studies focusing on diagnostic imaging (e.g., radiography, mammography, retinal imaging) or broader healthcare AI systems.	Articles without full-text access.
Articles proposing or evaluating fairness metrics like equalized odds, demographic parity, or predictive parity.	Studies that do not explicitly mention adversarial debiasing in their methodology.

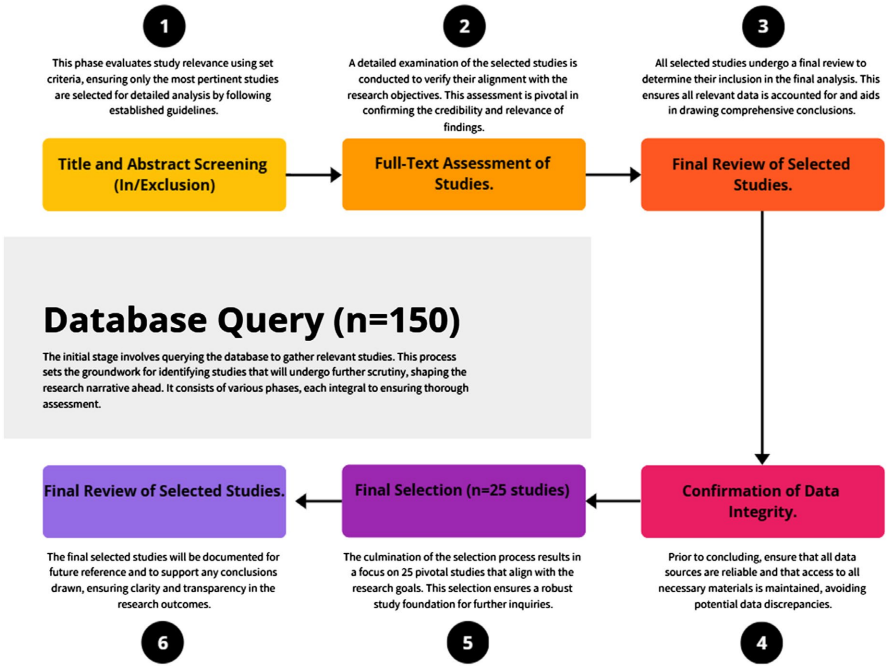

Figure 1. PRISMA Flowchart summarizing the study selection process.

3. Findings

Char *et al.* [8] examine the ethical challenges of applying machine learning in medicine. The authors mention that whereas AI can offer improved clinical outcomes, there exists an acute danger of biased judgments. Their work captures the need for effective ethical frameworks and cross-disciplinary surveillance. They conclude that technology must be preceded by effective governance and ethical counsel to ensure that AI improvements accrue to all patient groups on an equal basis.

In the groundbreaking paper that introduced Generative Adversarial Networks (GANs), Goodfellow *et al.* [9] presented the technical foundations upon which follow-on research in adversarial debiasing would proceed. Through its formulation of learning as a minimax game between a discriminator and a generator, the paper demonstrated the power of adversarial approaches for producing realistic data. Though not entirely focused on debiasing, the theoretical foundation they laid is now applied to fairness use in healthcare, and the authors are left to believe that adversarial models hold immense potential both for data synthesis and bias mitigation.

Mehrabi *et al.* [10] offer an extensive overview of bias and fairness in machine learning. They categorize bias into several forms that take place at various stages, from data acquisition to algorithm design, and address a range of ways for mitigating these biases, such as adversarial ones. Their overall finding is that while many approaches offer improvements, no single one is complete. They recommend more systematic approaches that integrate technical solutions (such as adversarial debiasing) with periodic audits and policy interventions.

Obermeyer *et al*. [11] provide an influential empirical analysis of racial bias in a health management algorithm. Their work documents how an algorithm designed to allocate healthcare resources systematically underestimated the needs of Black patients compared to White patients. The authors argue that incorporating fairness constraints is not only a technical enhancement but also a moral necessity. They conclude that rigorous audits and accountability standards must accompany algorithmic deployment in clinically sensitive domains.

Page *et al*. [7] set forth the PRISMA 2020 guidelines, which are essential for ensuring transparency and reproducibility in systematic reviews. Although this article is methodological rather than clinical, its advice is the foundation of our literature search and synthesis quality. The guidelines aim to increase review completeness and clarity, and their application is essential to any effort to review complex topics like healthcare adversarial debiasing.

Rajkomar *et al*. [12] stress the imperative need to ensure fairness in clinical algorithms. They illustrate how biased training data can lead to systemic disparity in AI predictions, with real implications for patient care. They point out that technical interventions like adversarial debiasing must be complemented with a change in data practices and regulations. They believe that ensuring health equity with AI will require algorithmic innovation as well as systemic reform in data collection.

Suresh and Guttag [13] propose a comprehensive framework to understand the various sources of harm that may occur in the machine learning life cycle. Their framework includes fairness and bias considerations, and they specifically refer to the instability issues common in adversarial training regimes. By identifying multiple failure modes, they conclude that continuous evaluation and adaptive methods are required to ensure that fairness-enhancing techniques do not compromise reliability.

By applying adversarial learning to bias mitigation, Zhang *et al*. [14] directly address the technical challenges of suppressing sensitive information in AI models. They demonstrate that incorporating an adversarial loss can significantly improve fairness metrics such as equalized odds, but also note that excessive penalization can lead to a reduction in overall accuracy. Their work highlights a central trade-off in debiasing approaches and calls for dynamic optimization strategies that balance fairness with clinical performance.

Chen *et al*. [15] explore whether AI has the potential to reduce disparities in medical and mental health care between groups of varying diversity. Their overview gives a balanced evaluation of the potential and limitations of AI-based interventions and the finding that biases within training material can seriously limit the potential of such systems. The authors encourage aggressive adoption of anti-bias strategies (such as adversarial debiasing) and state that although AI can potentially level the playing field, success will hinge on a rigorous, data-driven approach.

Esteva *et al*. [16] was a landmark for dermatology through the presentation of

deep neural networks' performance at classifying images of skin cancer, as well as experienced clinicians. Though their focus was on diagnostic performance, the study has broader implications regarding how bias would be introduced unless the training data are representative of the diverse population. Their result confirms the intuition that fairness in dataset curation is as important as technical debiasing methods applied downstream in the pipeline.

In diagnosing diabetic retinopathy, Gulshan *et al*. [17] showcased an AI system achieving high sensitivity and specificity compared to human experts. The work underscores the potential for AI to improve diagnostic accuracy, yet it also raises concerns about bias if population subgroups are underrepresented. Their conclusion, that AI can revolutionize screening practices, motivates subsequent efforts to incorporate adversarial debiasing to ensure equitable performance across demographic groups.

Arjovsky *et al*. [18] introduced the Wasserstein GAN, an advanced version of adversarial networks that significantly improves training stability. Although this paper focuses primarily on data generation, its conclusions about stabilizing adversarial models are extremely relevant to debiasing applications. The paper concludes that optimized adversarial training systems can counteract common flaws, thereby strengthening confidence in their application for fairness-related tasks.

In their exploration of "equality of opportunity" in supervised learning, Hardt *et al*. [19] propose that equitable classifiers should have similar true positive rates across all demographic subgroups. While not exclusively interested in adversarial methods, their formal definition of fairness conditions plays an important part in guiding follow-up technical solutions, including adversarial debiasing, to deliver fair predictions in healthcare systems. They determine that fairness constraints can be used in practice without excessive loss of overall performance.

Buolamwini & Gebru [20] conducted the seminal "Gender Shades" study, which documented significant biases in commercial facial recognition systems. According to their work, error rates differed greatly based on gender and skin color, prompting widespread calls for stricter fairness standards in AI development. Even though their work focuses primarily on computer vision, the implications have led to follow-on research in health care, related to the ethical imperatives of debiasing those algorithms used to control patient care. Their takeaway affirms that fairness must be a foundational consideration from data collection through the deployment of the models.

Chouldechova & Roth [21] present definitions of fairness in machine learning and also the trade-offs between them. Their survey is particularly helpful in the adversarial debiasing context as it also highlights that there isn't a single mathematical definition representing all aspects of fairness. Following this reasoning, the study concludes that applying fairness-enhancing techniques will necessarily rely on the presented context and fulfill the specific requirements of the system's clinical and operational requirements, so that applying adaptive adversarial frameworks becomes rational.

In their work, [22] criticize the shortcomings of oversimplified fairness metrics when applied to complex sociotechnical systems, contending that excessive dependence on general-purpose fairness formulations potentially conceals actual-world inequalities, a tendency to be referred to as "fairness gerrymandering." Drawing on their research, they conclude that continuous and context-aware evaluation needs to be carried out to guarantee that fairness interventions like adversarial debiasing accomplish their intended effects without inducing unintended harm.

Seminal work on fairness in machine learning, Barocas *et al*. [23] presents a comprehensive analysis of how and why bias happens and, as such, how it can be mitigated. Their book presents technical and policy insights because adversarial training is one among many tools required to treat systematic bias. They determine that the intersection of technological solutions and regulation will be essential to realizing fair AI deployment.

Kearns *et al*. [24] focus on the issue of subgroup fairness, ensuring that fairness is not only achieved in aggregate but also across smaller demographic groups. Their approach, that of auditing models for fairness gerrymandering, is supplemented by adversarial debiasing methods. The paper proves that strong mitigation of bias is a matter of cautious subpopulation performance and that adaptive auditing methods need to be incorporated into the development process.

Campanella *et al*. [25] demonstrate the potential of AI in computational pathology using a weakly supervised deep learning approach on whole slide images. Diagnostic-grade performance was achieved despite having fewer pixel-level annotations in their work. Although diagnostic accuracy was the primary focus, the study itself alludes to the responsible handling of heterogeneous data, thus necessitating the need for bias reduction techniques in pathological imaging.

Jiang *et al*. [26] provide a broad review of the evolution and applications of AI in healthcare. They touch on both the potential transformative impact of AI technologies and some of the challenges, such as data privacy, model interpretability, and bias. The review cautions that while AI has already begun to revolutionize clinical practice, the most important areas of concern, especially through avenues such as adversarial debiasing, will be tackling bias.

Kelly *et al*. [27] report on the key challenges that have to be addressed to meet the clinical promise of AI. The study alludes not just to technical ones such as interpretability and heterogeneity of the data, but to systemic challenges such as organizational inertia and uncertainty around the regulations. They conclude by calling for an interdisciplinary effort with measures such as adversarial debiasing, although very promising, having to be safely put within an array of checks and balances.

Wiens *et al*. [28] propose a comprehensive roadmap to ensure that machine learning systems in healthcare "do not harm." Their framework stresses the importance of incorporating safety, fairness, and ethical considerations at every stage of development. The paper concludes that despite significant advancements, the

integration of algorithms into clinical practice requires ongoing vigilance and interdisciplinary checks; adversarial debiasing is one critical piece of this complex puzzle.

Beam & Kohane [29] provide a balanced perspective on the promises and pitfalls of big data and machine learning in health care. They discuss how the integration of diverse data sources might enhance patient care, but also warn about issues related to privacy, data quality, and bias. Their study concludes that while innovations such as adversarial debiasing are promising, they must be applied with rigorous validation and ethical controls to prevent adverse effects.

In a landmark study comparing an AI algorithm (CheXNeXt) with radiologists for chest radiograph interpretation, Rajpurkar *et al*. [30] demonstrated that deep learning algorithms can achieve expert-level diagnostic performance. The study not only proved the clinical validity of the algorithm but also highlighted the need for bias evaluation as differences in performance between patient subgroups were noted. The authors conclude that future iterations of such systems will have to include bias mitigation measures to provide equal performance.

Topol [31] envisions a future of "high-performance medicine" where human ability and AI collaborate to transform healthcare delivery. Through a combination of multiple case studies, Topol describes how AI-driven systems can improve diagnostic accuracy, personalize treatment, and remove clinical inefficiencies. He concludes with cautionary hope that, if concerns regarding fairness, such as those addressed by adversarial debiasing, are surmounted, AI could be an excellent, equitable force in modern medicine.

4. Discussion

The discussed works provide a multi-faceted understanding of the promise and the challenges of debiasing adversaries in healthcare AI systems. At the most fundamental level, seminal works by [9] and [18] provide the underpinnings for building models that can generate realistic data distributions under the cover of concealing sensitive features. Such technology advancements have been used in medical environments using adversarial debiasing methods that directly address the devious issue of bias among clinical AI treatments. Based on the research we have examined, adversarial debiasing yields dramatic improvements in fairness metrics in radiological diagnosis imaging, and also in computed tomography (CT) imaging and chest radiograph interpretation. Rajpurkar *et al*. [30] also concur with improvements in fairness in chest radiograph interpretation and dermatologic imaging. While there are some trade-offs in sensitivity or overall performance, the evidence indicates that radiology is leading the way in taking advantage of adversarial debiasing techniques, and the techniques have potential for clinical practice.

One of the fundamental similarities across some studies is the inherent trade-off between fairness and predictive performance that is necessary. For example, [14] demonstrated that the application of adversarial debiasing achieved a considerable improvement in fairness metrics, such as equalized odds; however, this

came at the cost of reduced overall predictive performance and sensitivity. This trade-off presents one of the primary challenges: optimizing model fairness can, in certain instances, undermine essential clinical performance measures. This finding underscores the importance of carefully trading off bias removal with the maintenance of reliable diagnostic performance in clinical practice. In the same way, [32] have presented clear-cut evidence that adversarial debiasing is capable of inducing significant improvement in fairness measures, but at the cost of compromising accuracy or sensitivity. These trade-offs are particularly unwanted in diagnostic applications, where small decreases in accuracy can have life-or-death consequences. This tension between fairness and clinical efficacy means that any debiasing strategy will require a dynamic process to balance these two demands.

Apart from such technical concerns, several studies also highlight systemic and methodological concerns. Several of the papers covered here are reliant on retrospective datasets that may not capture all of the heterogeneity of clinical populations [17] [26]. Such a limitation highlights the need for prospective, multicenter trials that better determine the long-term efficacy and equity of adversarial debiasing methods. Moreover, the heterogeneity in network structures between CNNs and more recent transformer-based models complicates inter-study comparisons, necessitating a common platform for future research. [27] and [28] also mention that the technical benefits of debiasing interventions should be combined with regulatory standards, comprehensive audits, and ethical oversight.

There are also ethical and legal considerations emerging forcefully out of the study. [8] and [20] are a poignant reminder that technological solutions, adversarial debiasing in this instance, cannot ever be envisioned in siloed form. They must be placed in an overreaching ethical paradigm incorporating accountability, transparency, and equity at every level of development and deployment. The consensus across these studies is that while the technical solution proffered by adversarial debiasing seems appealing, truly making AI systems fair can happen only through the collaboration of the disciplines, between clinicians, policymakers, and technologists.

Finally, some authors call for the establishment of standardized standards and regulatory guidelines that are directed to the specific challenges of healthcare. For example, [21] and [23] suggest metrics that not only quantify bias at the aggregate level but also adjust for subgroup gaps. [22] and [24] also advocate for the use of fairness auditing techniques that can identify and rectify "fairness gerrymandering." Such an all-encompassing solution has the potential to revolutionize the existing landscape, ensuring that any improvement in predictive accuracy does not come at the cost of latent biases. Overall, the cumulative evidence from these studies indicates the promise of adversarial debiasing as a bias mitigation strategy for healthcare AI systems. But it also indicates that there is no silver bullet. The subtleties of clinical data, along with the delicate interplay between fairness and performance, require adaptive, context-dependent solutions that are constantly tested and iterated.

5. Conclusion and Future Directions

In short, adversarial debiasing is a leap forward in the quest to stem bias in healthcare AI systems. The literature reviewed herein demonstrates that adversarial techniques can effectively reduce the retention of sensitive information in machine learning models with improved fairness measures in a broad spectrum of applications, from diagnostic imaging to larger predictive analytics. Still, these improvements frequently come at a cost in accuracy, sensitivity, and model understandability, particularly when validations are done on past data sets.

In the future, this review will offer several avenues of research. First, future multicenter trials are urgently needed to determine the long-term clinical impact of adversarial debiasing methods. These studies would determine if the fairness improvements observed under laboratory conditions can be translated into better real-world patient outcomes. Second, developments of adaptive training protocols, such as reinforcement learning-based dynamic loss weighting, need to be explored to better balance accuracy and fairness without sacrificing important clinical performance.

Hybrid approaches that combine adversarial debiasing with ancillary bias minimization strategies like data augmentation or post hoc corrections are also promising and warrant full testing. However, these technological innovations must come with corresponding innovations in ethical and regulatory regimes. Developing formalized measures of fairness as well as audit procedures will foster trust and guarantee that AI systems used in the clinic deliver fair treatment.

Finally, interdisciplinary collaboration remains essential. All future efforts must facilitate increased dialogue between AI scientists, clinicians, ethicists, and policymakers to collaborate on the different dimensions of bias in medical AI. Collaborative approaches enable shared pools of high-quality diverse data to be created, context-dependent fairness metrics to be developed, and robust governance structures that ensure accountability throughout the AI life cycle can be created. In summary, while adversarial debiasing is no panacea, it is a worthwhile and promising step towards the achievement of equitable healthcare. By embracing technological innovation in combination with ethical, regulatory, and collaborative practices, the healthcare industry can unlock the revolutionary power of AI while preserving fairness, transparency, and the highest level of patient care.

Looking ahead, the rapid evolution of AI is poised to transform adversarial debiasing from an experimental technique into an integral, dynamic component of model training and deployment. Advances in self-supervised learning, dynamic optimization, and interpretability methods promise real-time adjustment of fairness constraints and continuous bias mitigation during inference, thereby balancing equity and diagnostic accuracy without significant human intervention. Moreover, hybrid approaches that combine adversarial debiasing with post hoc methods, reinforcement learning-based dynamic loss weighting, and meta-learning frameworks, exemplified in studies such as [6], are likely to extend these benefits into other clinical domains like oncology and personalized medicine. Ultimately,

as AI technologies progress, adversarial debiasing is expected to evolve into a continuously learning, context-aware mechanism, ensuring ethical and equitable AI deployment in healthcare and beyond.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Alowais, S.A., Alghamdi, S.S., Alsuhebany, N., Alqahtani, T., Alshaya, A.I., Almohareb, S.N., *et al.* (2023) Revolutionizing Healthcare: The Role of Artificial Intelligence in Clinical Practice. *BMC Medical Education*, **23**, Article No. 689. https://doi.org/10.1186/s12909-023-04698-z

[2] Secinaro, S., Calandra, D., Secinaro, A., Muthurangu, V. and Biancone, P. (2021) The Role of Artificial Intelligence in Healthcare: A Structured Literature Review. *BMC Medical Informatics and Decision Making*, **21**, Article No. 125. https://doi.org/10.1186/s12911-021-01488-9

[3] Hurlbert, M. (2025) Improving AI Performance for People of Color: Diagnosing Melanoma & Other Skin Cancers. Melanoma Research Alliance. https://www.curemelanoma.org/blog/making-ai-work-for-people-of-color-diagnosing-melanoma-and-other-skin-cancers

[4] Jemielity, S. (2025) Health Care Prediction Algorithm Biased against Black Patients, Study Finds. University of Chicago News. https://news.uchicago.edu/story/health-care-prediction-algorithm-biased-against-black-patients-study-finds

[5] Ramadass, S., Narayanan, S., Kumar, R. and K, T. (2024) Effectiveness of Generative Adversarial Networks in Denoising Medical Imaging (CT/MRI Images). *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-024-20130-0

[6] Zheng, G., Jacobs, M.A., Braverman, V. and Parekh, V.S. (2025) Towards Fair Medical AI: Adversarial Debiasing of 3D CT Foundation Embeddings. arXiv: 2502.04386. http://arxiv.org/abs/2502.04386

[7] Page, M.J., *et al.* (2021) The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ*, **372**, n71.

[8] Char, D.S., Shah, N.H. and Magnus, D. (2018) Implementing Machine Learning in Health Care—Addressing Ethical Challenges. *New England Journal of Medicine*, **378**, 981-983. https://doi.org/10.1056/nejmp1714229

[9] Goodfellow, I., Bengio, Y., Courville, A. and Bengio, Y. (2016) Deep Learning, Vol. 1, No. 2. MIT Press.

[10] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A. (2021) A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, **54**, 1-35. https://doi.org/10.1145/3457607

[11] Obermeyer, Z., Powers, B., Vogeli, C. and Mullainathan, S. (2019) Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science*, **366**, 447-453. https://doi.org/10.1126/science.aax2342

[12] Rajkomar, A., Hardt, M., Howell, M.D., Corrado, G. and Chin, M.H. (2018) Ensuring Fairness in Machine Learning to Advance Health Equity. *Annals of Internal Medicine*, **169**, 866-872. https://doi.org/10.7326/m18-1990

[13] Suresh, H. and Guttag, J. (2021) A Framework for Understanding Sources of Harm

Throughout the Machine Learning Life Cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 5-9 October 2021, 1-9. https://doi.org/10.1145/3465416.3483305

[14] Zhang, B.H., Lemoine, B. and Mitchell, M. (2018) Mitigating Unwanted Biases with Adversarial Learning. *Proceedings of the* 2018 *AAAI/ACM Conference on AI, Ethics, and Society*, New Orleans, 2-3 February 2018, 335-340. https://doi.org/10.1145/3278721.3278779

[15] Chen, I.Y., Szolovits, P. and Ghassemi, M. (2019) Can AI Help Reduce Disparities in General Medical and Mental Health Care? *AMA Journal of Ethics*, **21**, 167-179.

[16] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., *et al.* (2017) Dermatologist-level Classification of Skin Cancer with Deep Neural Networks. *Nature*, **542**, 115-118. https://doi.org/10.1038/nature21056

[17] Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., *et al.* (2016) Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, **316**, 2402-2410. https://doi.org/10.1001/jama.2016.17216

[18] Arjovsky, M., Chintala, S. and Bottou, L. (2017) Wasserstein Generative Adversarial Networks. *Proceedings of the* 34*th International Conference on Machine Learning*, Sydney, 6-11 August 2017, 214-223.

[19] Hardt, M., Price, E. and Srebro, N. (2016) Equality of Opportunity in Supervised Learning. *Proceedings of the* 30*th International Conference on Neural Information Processing Systems*, Barcelona, 5-10 December 2016, 3323-3331.

[20] Buolamwini, J. and Gebru, T. (2018) Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the* 1*st Conference on Fairness, Accountability and Transparency*, New York, 23-24 February 2018, 77-91.

[21] Chouldechova, A. and Roth, A. (2018) The Frontiers of Fairness in Machine Learning. arXiv: 1810.08810.

[22] Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S. and Vertesi, J. (2019) Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta, 29-31 January 2019, 59-68. https://doi.org/10.1145/3287560.3287598

[23] Barocas, S., Hardt, M. and Narayanan, A. (2023) Fairness and Machine Learning: Limitations and Opportunities. MIT Press.

[24] Kearns, M., Neel, S., Roth, A. and Wu, Z.S. (2018) Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. *Proceedings of the* 35*th International Conference on Machine Learning*, Stockholm, 10-15 July 2018, 2564-2572.

[25] Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., *et al.* (2019) Clinical-grade Computational Pathology Using Weakly Supervised Deep Learning on Whole Slide Images. *Nature Medicine*, **25**, 1301-1309. https://doi.org/10.1038/s41591-019-0508-1

[26] Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., *et al.* (2017) Artificial Intelligence in Healthcare: Past, Present and Future. *Stroke and Vascular Neurology*, **2**, 230-243. https://doi.org/10.1136/svn-2017-000101

[27] Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G. and King, D. (2019) Key Challenges for Delivering Clinical Impact with Artificial Intelligence. *BMC Medicine*, **17**, Article No. 195. https://doi.org/10.1186/s12916-019-1426-2

[28] Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V.X., Doshi-Velez, F., *et al.* (2019) Do No Harm: A Roadmap for Responsible Machine Learning for Health Care. *Nature*

Medicine, **25**, 1337-1340. https://doi.org/10.1038/s41591-019-0548-6

[29] Beam, A.L. and Kohane, I.S. (2018) Big Data and Machine Learning in Health Care. *JAMA*, **319**, 1317-1318. https://doi.org/10.1001/jama.2017.18391

[30] Rajpurkar, P., Irvin, J., Ball, R.L., Zhu, K., Yang, B., Mehta, H., *et al.* (2018) Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of the Chexnext Algorithm to Practicing Radiologists. *PLOS Medicine*, **15**, e1002686. https://doi.org/10.1371/journal.pmed.1002686

[31] Topol, E.J. (2019) High-Performance Medicine: The Convergence of Human and Artificial Intelligence. *Nature Medicine*, **25**, 44-56. https://doi.org/10.1038/s41591-018-0300-7

[32] Rajkomar, A., *et al.* (2018) Scalable and Accurate Deep Learning with Electronic Health Records. *npj Digital Medicine*, **1**, Article No. 18.