# Fairness Metrics in AI Healthcare Applications: A Review

Ibomoiye Domor Mienye
*Institute for Intelligent Systems*
*University of Johannesburg*
Johannesburg 2006, South Africa
ibomoiyem@uj.ac.za

Theo G. Swart
*Institute for Intelligent Systems*
*University of Johannesburg*
Johannesburg 2006, South Africa
tgswart@uj.ac.za

George Obaido
*Center for Human-Compatible Artificial Intelligence,*
*Berkeley Institute for Data Science,*
*University of California,*
*Berkeley, California, 94720, USA*
*gobaido@berkeley.edu*

***Abstract*—As artificial intelligence (AI) systems increasingly become popular in the healthcare sector, it is important to ensure the output of these technologies is fair and bias-free. This paper provides a concise survey of fairness metrics applied in healthcare AI, including their mathematical representations, suitable use cases, and limitations, which are lacking in the existing literature. The study also highlights the significance of implementing fairness metrics to ensure equitable outcomes across diverse patient populations and discusses the challenges and future directions in this rapidly evolving field.**

***Index Terms*—AI, bias, fairness metrics, healthcare, machine learning**

## I. Introduction

Artificial intelligence is transforming problem-solving and decision-making in various fields, including healthcare, finance, and transportation [1]–[3]. Machine learning (ML), a subset of AI, has the ability to process and analyze vast amounts of data at speeds that far surpass human capabilities. This has enabled researchers and professionals in various industries to gain valuable insights and make more informed decisions based on the data-driven predictions generated by ML models [4]–[6].

Furthermore, AI systems have achieved great success in the healthcare industry, particularly in the fields of medical imaging and diagnostics [7]–[9]. These AI systems can analyze medical images, such as X-rays and magnetic resonance imaging (MRI), with accuracy comparable to that of experienced radiologists. This speeds up the diagnostic process and significantly assists clinicians. Additionally, AI algorithms can analyze patient data to predict potential health risks and recommend personalized treatment plans, leading to improved patient outcomes [10]. However, concerns have been raised about the potential for bias in AI algorithms.

Bias and fairness in AI are important considerations that must be addressed in healthcare applications to ensure that AI systems are reliable, trustworthy, and equitable [11]. Fairness metrics in AI healthcare applications are crucial for evaluating the performance of AI algorithms and ensuring that they do not perpetuate existing biases in the training data or from the algorithm. These metrics can be used to assess the impact of AI algorithms on different demographic groups, such as race, gender, age, and socioeconomic status, to identify and mitigate potential biases.

Meanwhile, a major challenge in developing fairness metrics for AI healthcare applications is the lack of standardized definitions and methodologies [12], [13]. Different stakeholders may have varying interpretations of what constitutes fairness and bias, making it difficult to establish a universal set of metrics. Additionally, the complexity of healthcare data, which often includes sensitive information such as medical history and genetic data, poses unique challenges for developing fairness metrics that protect patient privacy while ensuring algorithmic transparency [14]. Despite these challenges, there are several promising approaches to improving fairness metrics in AI healthcare applications.

Therefore, this review explores the current state of fairness metrics in AI healthcare applications, including the challenges and opportunities for improvement. The review will lay the foundation for future research in developing standardized fairness metrics that can be applied across different AI healthcare applications to ensure equitable and unbiased decision-making processes.

The rest of this paper is organized as follows: Section II discusses some related works, and Section III presents a detailed background of the study, including the concept of fairness in AI, the importance of fairness, and AI applications in healthcare. Section IV discusses the different fairness metrics in AI, Section V presents some recommendations for implementing fairness metrics in AI applications in healthcare, and Section VI discusses challenges in developing fairness metrics. Lastly, Section VII concludes the study and presents future research directions.

## II. Related Work

The concept of fairness in AI, particularly in healthcare applications, has received significant attention from researchers and practitioners. In recent years, numerous studies have explored various aspects of fairness metrics, bias mitigation techniques, and ethical considerations in AI-driven healthcare systems. For example, Mhasawade et al. [15] studied the impact of AI algorithms on different demographic groups, such as race, gender, age, and socioeconomic status, to ensure equitable healthcare outcomes for all patients.

Similarly, Chin et al. [16] presented a framework for preventing bias in ML algorithms used for ML applications,

2835-5776/24/$31.00 ©2024 IEEE
DOI 10.1109/IRI62200.2024.00065

discussing the importance of fairness and equity. Furthermore, research in this area has highlighted the importance of transparency and interpretability in AI healthcare systems. For instance, Stiglic et al. [17] and Salahuddin et al. [18] investigated methods for making AI algorithms more transparent and understandable to patients, healthcare providers, and regulatory bodies.

Additionally, Peters et al. [19], Radanluev et al. [20], Karimian et al. [21] explored the ethical implications of using AI in healthcare and have proposed guidelines and frameworks for ensuring ethical and responsible AI deployment. These efforts emphasize the need for AI systems to uphold standards like beneficence, non-maleficence, autonomy, and justice in healthcare decision-making processes.

Overall, the related works show the complexity and importance of bias, ethics, and fairness in AI healthcare applications. Lastly, while there have been several research efforts on bias and fairness, including fairness metrics in general, there is limited research focused on developing and evaluating fairness metrics tailored specifically for healthcare applications. Therefore, building upon existing research, this study aims to contribute to the ongoing discourse on ethical and equitable AI deployment in healthcare, focusing on fairness metrics and providing valuable recommendations for researchers, policymakers, and healthcare practitioners. These metrics aim to quantify and address biases that may exist in AI algorithms used for medical diagnosis, treatment recommendation, and patient care.

## III. Background

This section provides an overview of the concept of fairness in AI, including the different types of biases that can exist in AI algorithms and the importance of addressing fairness in healthcare applications.

### A. Fairness in AI

Fairness is defined as the absence of discrimination or bias in decision-making processes. Ferrara [22] defined fairness in AI as the equitable treatment of individuals or groups, regardless of their demographic characteristics such as race, gender, age, or religion. Meanwhile, bias is the consistent and systematic deviation of a model's output from the true value. Bias in AI algorithms can arise from various sources, such as biased training data, algorithmic design, or human input. These biases can result in unfair outcomes, such as unequal access to healthcare services or inaccurate medical diagnoses.

Addressing fairness in AI is particularly crucial in healthcare applications, where decisions made by AI algorithms can have life-altering consequences for patients [23], [24]. For example, if a model used to predict disease risk is biased against certain demographic groups, it could lead to those individuals not receiving timely or appropriate medical care. This could worsen existing healthcare disparities and result in negative health outcomes for marginalized populations.

To ensure fairness in AI healthcare applications, it is essential to first identify and understand the types of biases that may exist in AI algorithms. This includes both explicit biases, which are intentionally programmed into the algorithm, and implicit biases, which are unintentional but still result in discriminatory outcomes [25]. Once these biases are identified, steps can be taken to mitigate their impact and ensure that AI algorithms are fair and equitable for all individuals, regardless of their demographic characteristics.

### B. Importance of Fairness Metrics in Healthcare AI

The integration of fairness metrics into AI healthcare applications is not just a technical necessity but a moral imperative. As AI technologies assume a more prominent role in healthcare decision-making, the need to ensure these systems operate equitably becomes paramount. This subsection explores the key reasons why fairness metrics are crucial in healthcare AI. Though not exhaustive, these reasons are outlined below:

*1) Promoting Equity in Healthcare:* Fairness metrics serve as a crucial tool in identifying and mitigating biases in AI systems, thereby promoting equity [26]. They ensure that AI applications do not perpetuate existing healthcare disparities or introduce new forms of discrimination.

*2) Enhancing Patient Trust and Engagement:* Patients are more likely to trust and engage with AI-driven healthcare services when they are assured of fair and unbiased treatment [27], [28]. Fairness metrics can help build this trust by ensuring that AI applications treat all patients equitably, regardless of their background.

*3) Improving Healthcare Outcomes for All:* Fairness metrics contribute to better healthcare outcomes across diverse patient populations by ensuring that AI systems are fair and equitable [29]. This is particularly important in global health scenarios, where disparities can be pronounced.

*4) Economic Efficiency:* Fair and unbiased AI systems can lead to more efficient allocation of healthcare resources, reducing wasteful expenditures and ensuring that interventions are directed where they are most needed.

*5) Legal and Ethical Compliance:* Incorporating fairness metrics can assist healthcare organizations in complying with legal and ethical standards. Regulations increasingly require that AI systems, especially those in sensitive areas like healthcare, operate transparently and equitably, making fairness metrics indispensable [30]–[32].

### C. AI Applications in Healthcare

AI has been applied in different areas within the healthcare space. This section discusses some of those applications. Firstly, a popular application of AI in healthcare is medical imaging. AI algorithms can be used to analyze complex medical images, such as MRIs and CT scans, to detect hidden abnormalities that human radiologists may miss. This can help to improve the accuracy of diagnoses and reduce the likelihood of misdiagnosis [33]. In addition, AI-based imaging systems can also assist in surgical procedures by providing surgeons with real-time feedback, helping them make more precise incisions and reduce the risk of complications [34].

Another important application of AI in healthcare is in diagnostics. AI algorithms can analyze vast amounts of patient data, including medical history, lab results, and genetic information, to help healthcare providers make more accurate diagnoses [35]. Also, machine learning models have been developed to predict potential health risks in patients and their predisposition to certain diseases, leading to early detection and intervention. Personalized medicine is another area where AI is making a significant impact. ML algorithms can help healthcare providers tailor treatment plans to individual patients based on their unique characteristics, such as genetic makeup, lifestyle factors, and medical history [36], [37]. This personalized approach to medicine can lead to more effective treatments, fewer side effects, and better patient outcomes.

Predictive analytics is another key application of AI in healthcare. ML models can identify trends and patterns that can help healthcare providers predict future health events. This can be useful in managing chronic diseases, such as diabetes, where early intervention can help prevent complications and improve patient outcomes. Thereby aiding healthcare providers in identifying high-risk patients and providing targeted interventions to prevent disease progression, which can lead to better health outcomes and reduced healthcare costs [38], [39].

## IV. Fairness Metrics

This section defines and describes the various fairness metrics, including their mathematical formulations. We also discuss the classification of fairness metrics. Meanwhile, fairness metrics are quantitative measures designed to assess the extent to which AI models adhere to the principles of equity and justice [13]. The various fairness metrics include:

*1) Group Fairness:* Group fairness, also known as statistical or demographic fairness, requires that predictive outcomes are independent of specific sensitive attributes such as race, gender, or age [40]. Assuming $\hat{Y}$ is the predicted outcome and $D$ is the sensitive attribute with a given value $d$, then group fairness can be expressed mathematically as:

$$P(\hat{Y} = 1|D = d) = P(\hat{Y} = 1|D \neq d) \tag{1}$$

This formulation ensures that the probability of a positive outcome, for instance, is the same across different groups defined by $D$. Furthermore, group fairness aims to equalize some statistical measure, such as positive predictive value (PPV), false discovery rate (FDR), etc, across groups defined by the sensitive attribute. For example, ensuring equal FDR across groups can be critical in applications where the cost of false positives is high:

$$FDR_{group_1} = FDR_{group_2} = \cdots = FDR_{group_n} \tag{2}$$

*2) Individual Fairness:* Individual fairness mandates that similar individuals should receive similar predictions. It can be mathematically formalized using a distance metric $d(\cdot)$, which measures the similarity between individuals:

$$d(\hat{Y}_i, \hat{Y}_j) \leq D(x_i, x_j) \quad \forall i, j \tag{3}$$

where $x_i$ and $x_j$ are feature vectors of individuals $i$ and $j$, $\hat{Y}_i$ and $\hat{Y}_j$ are the predicted outcomes, and $D(\cdot)$ is a metric that quantifies the difference in treatment or outcomes [41], [42]. In healthcare applications, this might involve considering not just direct feature similarities but also underlying health conditions and other medically relevant factors that influence health outcomes. Therefore, a more nuanced approach to individual fairness might use a weighted distance metric that accounts for the relative importance of different features:

$$d(\hat{Y}_i, \hat{Y}_j) \leq w \cdot D(x_i, x_j) \quad \forall i, j \tag{4}$$

where $w$ represents a set of weights indicating the importance of each feature in $x_i$ and $x_j$ in determining treatment fairness.

*3) Equality of Opportunity:* Equality of opportunity is an extension of group fairness, which states that groups should have equal true positive rates. This is crucial in healthcare, where equal access to treatment opportunities is essential [43]. This metric is defined as:

$$P(\hat{Y} = 1|D = d, Y = 1) = P(\hat{Y} = 1|D \neq d, Y = 1) \tag{5}$$

where $Y$ represents the true outcome. This condition ensures that, for individuals who should receive a positive outcome, the chance of being correctly identified is equal across different groups [43].

*4) Demographic Parity:* Demographic parity, also known as statistical parity, is a fairness metric that requires the decision outcomes to be independent of the sensitive attributes. It aims for the probability of a positive prediction to be the same across different groups defined by the sensitive attribute [44]. Mathematically, it is defined as:

$$P(\hat{Y} = 1|D = d) = P(\hat{Y} = 1|D \neq d) \tag{6}$$

This definition is similar to that of group fairness but emphasizes the need for equal representation across outcomes, irrespective of the sensitive attribute's distribution within the population.

*5) Equalized Odds:* Equalized odds extend the concept of fairness by requiring that the model's accuracy, (both true positive rate and false positive rate) be the same across groups. This metric is particularly relevant in scenarios where both types of errors (i.e., false positives and false negatives) are critical [45]. Equalized odds can be expressed as:

$$P(\hat{Y} = 1|D = d, Y = y) = P(\hat{Y} = 1|D \neq d, Y = y) \quad \forall y \in \{0, 1\} \tag{7}$$

This ensures that the model is equally accurate for all groups, regardless of the actual outcome $Y$. Equalized odds ensure a balanced approach to fairness, where the AI model's accuracy and error rates do not disproportionately affect any group, aligning with more comprehensive fairness objectives.

*6) Counterfactual Fairness:* Counterfactual fairness considers a model to be fair if its predictions are the same in the actual world and a counterfactual world where the sensitive attribute $D$ is different, but everything else remains the same [46]. A model is counterfactually fair if:

$$P(\hat{Y}_{D\leftarrow d}(U) = y|X = x, D = d) = P(\hat{Y}_{D\leftarrow d'}(U) = y|X = x, D = d) \tag{8}$$

for all $y$, where $X$ are non-sensitive attributes, $D$ is the sensitive attribute, $d$ and $d'$ are possible values of $D$, and $U$ represents the underlying factors that influence both $X$ and $Y$. This metric addresses the causal relationships between attributes and outcomes, ensuring fairness in a more comprehensive manner.

## V. Recommendations for Implementing Fairness Metrics in Healthcare AI

The application of fairness metrics in healthcare AI is important for identifying and mitigating biases that may lead to disparities in patient care. This section recommends how various fairness metrics can be applied to different aspects of healthcare AI, from diagnosis to treatment recommendations and patient management.

### A. Diagnosis and Treatment Recommendations

In diagnostic AI applications, especially ML models used in detecting diseases, fairness metrics can ensure that the model's performance is consistent across groups defined by sensitive attributes. A good example will be using the equality of opportunity metric, which can help ensure that models for diagnosing diseases do not disproportionately misclassify patients from certain racial or ethnic backgrounds. Mathematically, this involves evaluating the true positive rate equality:

$$TPR_{group_1} = TPR_{group_2} = \cdots = TPR_{group_n} \tag{9}$$

where $TPR_{group_i}$ is the true positive rate for group $i$. Achieving similar TPRs across groups indicates that the model offers equitable diagnostic accuracy.

### B. Resource Allocation and Patient Scheduling

Fairness in resource allocation and patient scheduling AI-based systems can be assessed using group fairness metrics to ensure that resources, such as appointment slots or access to specialized care, are distributed equitably among patients, regardless of demographic factors. This application requires continuous monitoring and adjustment of the models to maintain fairness over time, especially in dynamic environments like hospitals where patient demographics and resource availability may change.

### C. Personalized Medicine

Individual fairness metrics are well-suited for personalized medicine applications, where treatment plans are tailored to the individual characteristics of patients. Thereby ensuring that similar patients receive similar treatment recommendations. This requires ML developers to have an understanding of individual fairness and ensure models are trained using algorithms that consider the unique medical histories and conditions of patients. This can be represented mathematically as:

$$d(\hat{Y}_i, \hat{Y}_j) \leq \epsilon \quad \text{for similar } x_i, x_j \tag{10}$$

$d(\cdot)$ is a function measuring the discrepancy in treatment recommendations, and $\epsilon$ is a small tolerance level, ensuring that the recommendations are fair for patients with similar profiles.

### D. Monitoring and Evaluation

Continuous monitoring and evaluation of ML models in healthcare using fairness metrics are essential to ensure long-term equity in patient outcomes. This involves the initial application of fairness metrics and the regular reassessment of the models as new data become available and healthcare practices evolve. Tools and frameworks for continuous fairness assessment can assist in adapting ML models to maintain fairness standards over time.

## VI. Challenges in Developing Fairness Metrics

The development of fairness metrics for AI in healthcare applications poses several challenges, spanning technical, ethical, and practical domains. Addressing these challenges is crucial to ensure the equitable and unbiased deployment of AI technologies in healthcare settings.

### A. Data Representation and Bias

One of the primary challenges in developing fairness metrics lies in addressing biases present in the data used to train the models [47]. Healthcare data often reflect societal biases and disparities, leading to biased models that may perpetuate existing inequalities. Overcoming these biases requires careful attention to data collection, labelling, and processing practices to ensure that AI systems are trained on diverse and representative datasets [48], [49].

### B. Complexity in Defining Fairness

Fairness is a multifaceted and subjective concept that can vary across cultures, contexts, and stakeholders [48]. Defining fairness in the context of AI healthcare applications is a significant challenge, as different stakeholders may have divergent interpretations of what constitutes fair treatment. Developing universally acceptable definitions of fairness and bias that account for diverse perspectives and values is essential for the effective implementation of fairness metrics.

### C. Trade-offs Between Fairness and Model Performance

Balancing fairness with other performance metrics, such as accuracy and efficiency, presents a considerable challenge in developing fairness metrics for AI healthcare applications [50]. Optimizing AI models for fairness often involves trade-offs that may impact overall model performance. Finding the optimal balance between fairness and performance requires careful consideration of the specific healthcare context and stakeholders' priorities.

### D. Lack of Standardization and Regulation

The lack of standardized definitions, methodologies, and regulatory frameworks for fairness metrics in AI healthcare applications hinders progress in this field. Without clear guidelines and regulations, developers may struggle to consistently implement and evaluate fairness metrics across different AI systems and applications [14], [22]. Establishing standardized frameworks and regulatory guidelines is essential for ensuring consistency and accountability in the development and deployment of fairness metrics in healthcare AI.

### E. Monitoring and Maintaining Fairness Over Time

Ensuring fairness in AI healthcare applications is an ongoing process that requires continuous monitoring and adaptation. AI systems are dynamic and may evolve over time, which causes challenges in maintaining fairness as algorithms interact with new data and environments [51]–[53]. Developing mechanisms for the ongoing evaluation and adjustment of fairness metrics is essential for ensuring that AI systems remain fair and equitable throughout their lifecycle.

## VII. Conclusion and Future Research directions

This study presented a concise survey of fairness metrics in AI, focusing on applications in the healthcare space. The study provided a comprehensive definition and the applications of the various fairness metrics, including group fairness, individual fairness, demographic parity, equality of opportunity, equalized odds, and counterfactual fairness. It was established that fairness metrics in AI healthcare applications are essential for ensuring that the output of AI models is fair, unbiased, and equitable for all individuals, regardless of their demographic characteristics.

Future research can focus on developing standardized definitions and methodologies for measuring bias, implementing transparency measures to ensure algorithmic accountability, and addressing data bias in AI algorithms. By taking these steps, researchers can help advance the use of fairness metrics in AI healthcare applications and ensure that AI technologies are used responsibly and ethically to benefit patients and healthcare providers. Additionally, future research can focus on developing new fairness metrics that can account for the unique challenges of healthcare data, such as patient privacy and data sensitivity. Addressing these challenges can advance the field of fairness metrics in AI healthcare applications and ensure that AI technologies are used in a responsible and ethical manner.

## References


[1] J. Guan, "Artificial intelligence in healthcare and medicine: Promises, ethical challenges and governance," *Chinese Medical Sciences Journal*, vol. 34, no. 2, pp. 76–83, 2019.

[2] H. H. Al-Baity, "The artificial intelligence revolution in digital finance in saudi arabia: A comprehensive review and proposed framework," *Sustainability*, vol. 15, no. 18, 2023.

[3] T. O'Halloran, G. Obaido, B. Otegbade, and I. D. Mienye, "A deep learning approach for maize lethal necrosis and maize streak virus disease detection," *Machine Learning with Applications*, vol. 16, p. 100556, 6 2024.

[4] E. Kyrimi, S. McLachlan, K. Dube, M. R. Neves, A. Fahmi, and N. Fenton, "A comprehensive scoping review of Bayesian networks in healthcare: Past, present and future," *Artificial Intelligence in Medicine*, vol. 117, p. 102108, 2021.

[5] G. Obaido, B. Ogbuokiri, C. W. Chukwu, F. J. Osaye, O. F. Egbelowo, M. I. Uzochukwu, I. D. Mienye, K. Aruleba, M. Primus, and O. Achilonu, "An improved ensemble method for predicting hyperchloremia in adults with diabetic ketoacidosis," *IEEE Access*, 2024.

[6] I. D. Mienye and Y. Sun, "A machine learning method with hybrid feature selection for improved credit card fraud detection," *Applied Sciences*, vol. 13, no. 12, p. 7254, 2023.

[7] L. Nanni, S. Brahnam, M. Paci, and S. Ghidoni, "Comparison of different convolutional neural network activation functions and methods for building ensembles for small to midsize medical data sets," *Sensors*, vol. 22, no. 16, p. 6129, 2022.

[8] S. H. Yoo, H. Geng, T. L. Chiu, S. K. Yu, D. C. Cho, J. Heo, M. S. Choi, I. H. Choi, C. Cung Van, N. V. Nhung, B. J. Min, and H. Lee, "Deep learning-based decision-tree classifier for covid-19 diagnosis from chest x-ray imaging," *Frontiers in Medicine*, vol. 7, 7 2020.

[9] I. D. Mienye, P. Kenneth Ainah, I. D. Emmanuel, and E. Esenogho, "Sparse noise minimization in image classification using genetic algorithm and densenet," in *2021 Conference on Information Communications Technology and Society (ICTAS)*, pp. 103–108, 2021.

[10] I. D. Mienye and Y. Sun, "Effective feature selection for improved prediction of heart disease," in *Pan-African Artificial Intelligence and Smart Systems* (T. M. N. Ngatched and I. Woungang, eds.), (Cham), pp. 94–107, Springer International Publishing, 2022.

[11] Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, and O. Russakovsky, "Towards fairness in visual recognition: Effective strategies for bias mitigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[12] M. Madaio, L. Egede, H. Subramonyam, J. Wortman Vaughan, and H. Wallach, "Assessing the fairness of ai systems: Ai practitioners' processes, challenges, and needs for support," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, pp. 1–26, 3 2022.

[13] A. Agarwal, H. Agarwal, and N. Agarwal, "Fairness score and process standardization: framework for fairness certification in artificial intelligence systems," *AI and Ethics*, vol. 3, pp. 267–279, 3 2022.

[14] S. Caton and C. Haas, "Fairness in machine learning: A survey," *ACM Computing Surveys*, 2020.

[15] V. Mhasawade, Y. Zhao, and R. Chunara, "Machine learning and algorithmic fairness in public and population health," *Nature Machine Intelligence*, vol. 3, no. 8, pp. 659–666, 2021.

[16] M. H. Chin, N. Afsar-Manesh, A. S. Bierman, C. Chang, C. J. Colón-Rodríguez, P. Dullabh, D. G. Duran, M. Fair, T. Hernandez-Boussard, M. Hightower, *et al.*, "Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care," *JAMA Network Open*, vol. 6, no. 12, pp. e2345050–e2345050, 2023.

[17] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, "Interpretability of machine learning-based prediction models in healthcare," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 5, p. e1379, 2020.

[18] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, and P. Lambin, "Transparency of deep neural networks for medical image analysis: A review of interpretability methods," *Computers in biology and medicine*, vol. 140, p. 105111, 2022.

[19] D. Peters, K. Vold, D. Robinson, and R. A. Calvo, "Responsible ai—two frameworks for ethical design practice," *IEEE Transactions on Technology and Society*, vol. 1, no. 1, pp. 34–47, 2020.

[20] P. Radanliev, O. Santos, A. Brandon-Jones, and A. Joinson, "Ethics and responsible ai deployment," *Frontiers in Artificial Intelligence*, vol. 7, p. 1377011, 2024.

[21] G. Karimian, E. Petelos, and S. M. Evers, "The ethical issues of the application of artificial intelligence in healthcare: a systematic scoping review," *AI and Ethics*, vol. 2, no. 4, pp. 539–551, 2022.

[22] E. Ferrara, "Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies," *Sci*, vol. 6, no. 1, 2024.

[23] D. Cirillo, S. Catuara-Solarz, C. Morey, E. Guney, L. Subirats, S. Mellino, A. Gigante, A. Valencia, M. J. Rementeria, A. S. Chadha, and N. Mavridis, "Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare," *npj Digital Medicine*, vol. 3, 6 2020.

[24] P. Mosteiro, J. Kuiper, J. Masthoff, F. Scheepers, and M. Spruit, "Bias discovery in machine learning models for mental health," *Information*, vol. 13, no. 5, 2022.

[25] M. DeCamp and C. Lindvall, "Latent bias and the implementation of artificial intelligence in medicine," *Journal of the American Medical Informatics Association*, vol. 27, pp. 2020–2023, 6 2020.

[26] T. P. Pagano, R. B. Loureiro, F. V. N. Lisboa, R. M. Peixoto, G. A. S. Guimarães, G. O. R. Cruz, M. M. Araujo, L. L. Santos, M. A. S. Cruz, E. L. S. Oliveira, I. Winkler, and E. G. S. Nascimento, "Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods," *Big Data and Cognitive Computing*, vol. 7, no. 1, 2023.

[27] W. A. Rogers, H. Draper, and S. M. Carter, "Evaluation of artificial intelligence clinical applications: Detailed case analyses show value of healthcare ethics approach in identifying patient care issues," *Bioethics*, vol. 35, pp. 623–633, 5 2021.

[28] D. Dreesens, A. Stiggelbout, T. Agoritsas, G. Elwyn, S. Flottorp, J. Grimshaw, L. Kremer, N. Santesso, D. Stacey, S. Treweek, M. Armstrong, A. Gagliardi, S. Hill, F. Légaré, R. Ryan, P. Vandvik, and T. van der Weijden, "A conceptual framework for patient-directed knowledge tools to support patient-centred care: Results from an evidence-informed consensus meeting," *Patient Education and Counseling*, vol. 102, no. 10, pp. 1898–1904, 2019.

[29] M. Liu, Y. Ning, S. Teixayavong, M. Mertens, J. Xu, D. S. W. Ting, L. T.-E. Cheng, J. C. L. Ong, Z. L. Teo, T. F. Tan, *et al.*, "A translational perspective towards clinical ai fairness," *NPJ Digital Medicine*, vol. 6, no. 1, p. 172, 2023.

[30] P. G. R. de Almeida, C. D. dos Santos, and J. S. Farias, "Artificial intelligence regulation: a framework for governance," *Ethics and Information Technology*, vol. 23, pp. 505–525, 4 2021.

[31] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature Machine Intelligence*, vol. 1, pp. 389–399, 9 2019.

[32] N. Balasubramaniam, M. Kauppinen, K. Hiekkanen, and S. Kujala, "Transparency and explainability of ai systems: Ethical guidelines in practice," in *Requirements Engineering: Foundation for Software Quality* (V. Gervasi and A. Vogelsang, eds.), (Cham), pp. 3–18, Springer International Publishing, 2022.

[33] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Informatics in Medicine Unlocked*, vol. 20, p. 100402, 2020.

[34] L. Xu, H. Zhang, J. Wang, A. Li, S. Song, H. Ren, L. Qi, J. J. Gu, and M. Q.-H. Meng, "Information loss challenges in surgical navigation systems: From information fusion to ai-based approaches," *Information Fusion*, vol. 92, pp. 13–36, 2023.

[35] I. D. Mienye, G. Obaido, K. Aruleba, and O. A. Dada, "Enhanced prediction of chronic kidney disease using feature selection and boosted classifiers," in *International Conference on Intelligent Systems Design and Applications*, pp. 527–537, Springer, 2021.

[36] M. Sebastiani, C. Vacchi, A. Manfredi, and G. Cassone, "Personalized medicine and machine learning: A roadmap for the future," *Journal of Clinical Medicine*, vol. 11, no. 14, 2022.

[37] S. Vadapalli, H. Abdelhalim, S. Zeeshan, and Z. Ahmed, "Artificial intelligence and machine learning approaches using gene expression and variant data for personalized medicine," *Briefings in Bioinformatics*, vol. 23, 5 2022.

[38] J. K. Silver, "Cancer prehabilitation and its role in improving health outcomes and reducing health care costs," *Seminars in Oncology Nursing*, vol. 31, no. 1, pp. 13–30, 2015. Emerging Issues in Cancer.

[39] I. D. Mienye and Y. Sun, "Heart disease prediction using enhanced machine learning techniques," in *Intelligent Systems and Machine Learning for Industry*, pp. 93–114, CRC Press, 2022.

[40] D. Pessach and E. Shmueli, "Algorithmic fairness," in *Machine Learning for Data Science Handbook*, pp. 867–886, Springer International Publishing, 2023.

[41] P. George John, D. Vijaykeerthy, and D. Saha, "Verifying individual fairness in machine learning models," in *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)* (J. Peters and D. Sontag, eds.), vol. 124 of *Proceedings of Machine Learning Research*, pp. 749–758, PMLR, 03–06 Aug 2020.

[42] S. Sharifi-Malvajerdi, M. Kearns, and A. Roth, "Average individual fairness: Algorithms, generalization and experiments," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[43] J. E. Roemer and A. Trannoy, "Chapter 4 - equality of opportunity," in *Handbook of Income Distribution* (A. B. Atkinson and F. Bourguignon, eds.), vol. 2 of *Handbook of Income Distribution*, pp. 217–300, Elsevier, 2015.

[44] A. Pereira Barata, F. W. Takes, H. J. van den Herik, and C. J. Veenman, "Fair tree classifier using strong demographic parity," *Machine Learning*, 8 2023.

[45] Y. Romano, S. Bates, and E. Candes, "Achieving equalized odds by resampling sensitive attributes," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 361–371, Curran Associates, Inc., 2020.

[46] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[47] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[48] R. Schwartz, R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, and P. Hall, *Towards a standard for identifying and managing bias in artificial intelligence*, vol. 3. US Department of Commerce, National Institute of Standards and Technology, 2022.

[49] T. P. Pagano, R. B. Loureiro, F. V. Lisboa, R. M. Peixoto, G. A. Guimarães, G. O. Cruz, M. M. Araujo, L. L. Santos, M. A. Cruz, E. L. Oliveira, *et al.*, "Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods," *Big data and cognitive computing*, vol. 7, no. 1, p. 15, 2023.

[50] J. S. Kim, J. Chen, and A. Talwalkar, "Fact: A diagnostic for group fairness trade-offs," in *International Conference on Machine Learning*, pp. 5264–5274, PMLR, 2020.

[51] P. Esmaeilzadeh, "Challenges and strategies for wide-scale artificial intelligence (ai) deployment in healthcare practices: A perspective for healthcare organizations," *Artificial Intelligence in Medicine*, p. 102861, 2024.

[52] D. Leslie, C. Rincon, M. Briggs, A. Perini, S. Jayadeva, A. Borda, S. Bennett, C. Burr, M. Aitken, M. Katell, *et al.*, "Ai fairness in practice," *arXiv preprint arXiv:2403.14636*, 2024.

[53] P. Chen, L. Wu, and L. Wang, "Ai fairness in data management and analytics: A review on challenges, methodologies and applications," *Applied sciences*, vol. 13, no. 18, p. 10258, 2023.