ISSN 2664-4150 (Print) & ISSN 2664-794X (Online)
South Asian Research Journal of Engineering and Technology
Abbreviated Key Title: South Asian Res J Eng Tech

| Volume-7 | Issue-1 | Jan-Feb- 2025 | **DOI: https://doi.org/10.36346/sarjet.2025.v07i01.004**

Original Research Article

Developing Frameworks for Assessing and Mitigating Bias in AI Systems: A Case Study on Ensuring Fairness in AI Diagnostic Tools through Diverse Training Datasets to Prevent Misdiagnosis in Underrepresented Populations

Opeyemi Oluwagbenga Owolabi[1*], Opeyemi Bilqees Adewusi[1], Funmilayo Arinola Ajayi[1], Ajoke A. Asunmonu[2], Ozoemezim Chukwurimazu[3], Jefferson Ederhion[4], Ohi Moses Ayeni[5]

[1]Member, the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB)
[2]University of Benin
[3]University of Florida
[4]University of Maryland, College Park
[5]University of Swansea

***Corresponding Author:** Opeyemi Oluwagbenga Owolabi
Member, the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB)

Article History
Received: 20.01.2025
Accepted: 25.02.2025
Published: 27.02.2025

Abstract: Indeed, AI on the rise in health has given a way forward to cause revolutionary changes in diagnostic capability. With this, critical issues relating to bias and fairness arise, particularly regarding the under-representative in any given dataset. It attempts to explore the relevance of probably the leading challenge that biases in diagnostic artificial intelligence impose on health and the contribution of representative, various train datasets toward reducing such biases for guaranteeing health equity. This study is a quantitative research in which data through purposive sampling was collected with the use of survey methodology from 160 stakeholders, including clinicians, patients, and AI developers. The structured questionnaires captured perceptions related to AI bias, fairness, and the effectiveness of mitigation strategies. Results findings show great concerns over biases in AI systems, with 48.75% agreeing that bias within the training dataset is a major hindrance in clinical decision-making. A majority of 53.33% said it is the diversity in the datasets that helps in promoting fairness, while 76.88% support frequent algorithmic audits. Other key strategies recommended were collaboration among stakeholders and transparency. The current study, therefore, infers that these biases in diagnostic Artificial Intelligence tools need to be treated along the dimensions of data diversity, auditing algorithms, and collaboration with various stakeholders, to whom the transparency of processing should be allowed. Hence, there is a need for improvement in robust frameworks assurance for the equity of AI so that it could guarantee fairness, reliability, and inclusion among various populations for diagnoses.

Keywords: AI Bias, Fairness in Machine Learning, Bias Mitigation Strategies, Diverse Training Datasets, Healthcare AI, Underrepresented Populations.

1.0 INTRODUCTION

AI has revolutionised diagnoses and treatments and further personalised results in many areas, such as cardiology, ophthalmology, and dermatology [1]. There are two ways, in which AI can outperform human performance in health; First, AI can learn from big data that no clinician could ever learn from, and a well-designed AI system will extract the necessary information efficiently from offline or online data that would assist in improving institutional performance and helping the health professional arrive at informed decisions in problem-solving time; second, AI systems can perform predefined tasks with high precision, and can be able to work actively without performance compromise in operation, never getting burnt out from work; this feature of AI technology could change the face of complicated surgeries [2]. The importance of AI can be observed in Fig 1 which illustrates its benefits in healthcare.

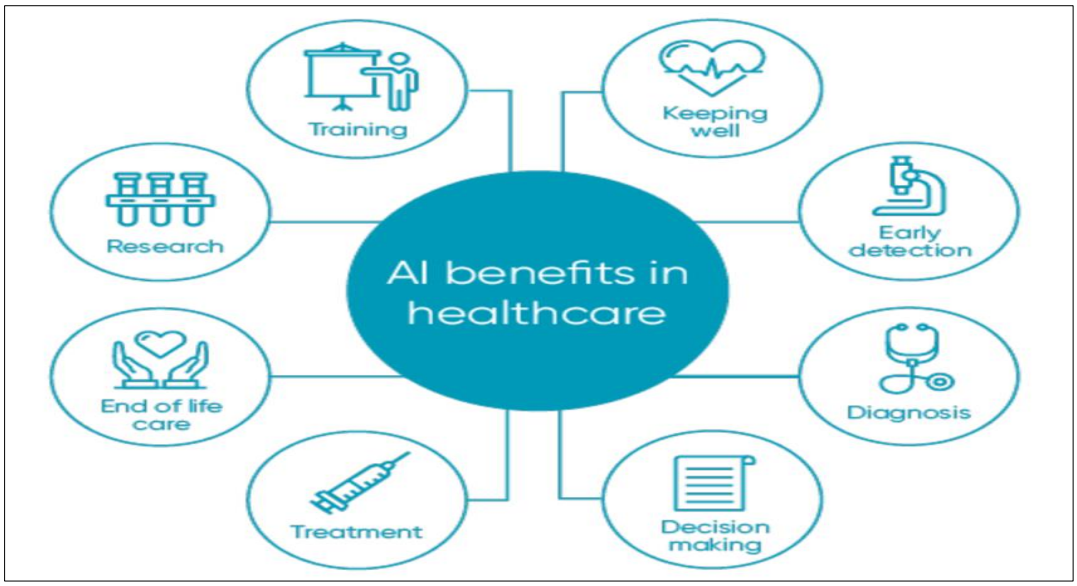

Figure 1: Importance of AI in healthcare [3]

This development, however, is accompanied by several challenges that AI systems in health pose to bias in care delivery, accuracy of diagnosis, and treatment outcome across demographic groups [1-4]. One of the major challenges within medical AI has to do with imbalanced population data, where generally biased algorithms tend to affect historically disparaged groups; this neglection of population data in the composition of training datasets makes biases in algorithms wrong, thus misdiagnosing and adding to healthcare incongruity; that would mean the model learns only the patterns and features pertinent to the demographic or diagnostic group if the training dataset is dominantly from that particular group; a direct implication of such facts is poor performance in diverse populations, or rather biased predictions since the model would have learned inadequately about the under-represented groups, and such a model of diagnostic AI, when developed in a clinical setting from imbalanced datasets, is overly obsessed with conditions characteristic of the most represented group; this would later cause under-exploration or misidentification to happen due to the conditions being not so prevailing, represented dissimilarly among groups and few in numbers [5], as seen in figure 2.

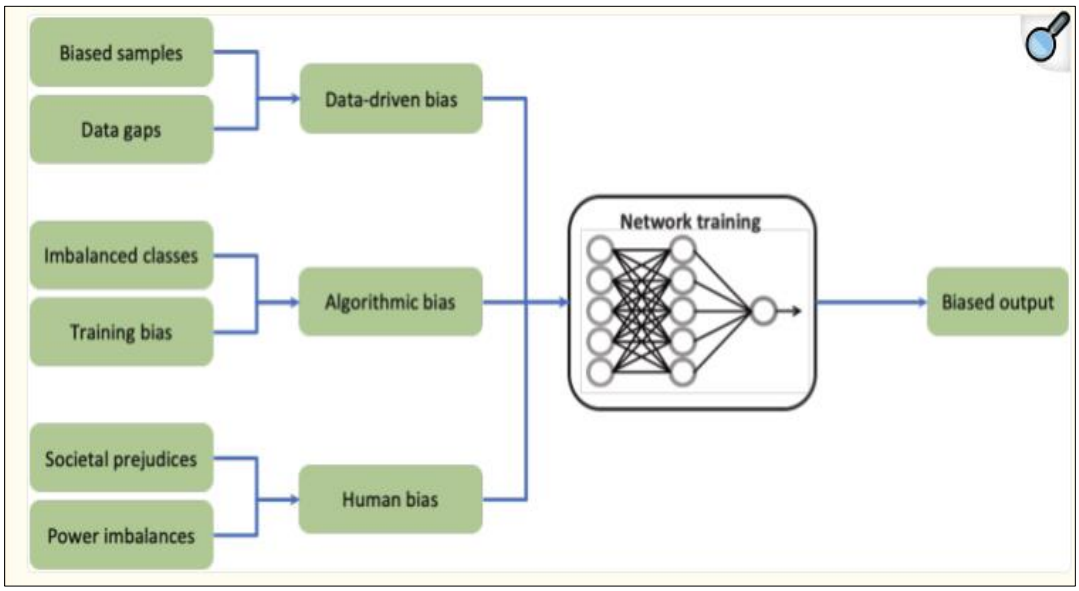

Figure 2: The different sources of bias in machine learning algorithms training datasets [6]

Also, some bias could crop up anywhere along the pipeline in data collection through development to model completion, all to come down on users and organisations deploying those diagnostic AI systems [7, 8]. This might result in poor performance of the models on various subgroups of patients, rendering the predictions clinically not meaningful, therefore further increasing existing health disparities [7-9]. For example, the performance of an AI algorithm trained almost exclusively on the data of white patients is, without doubt, poorly performing for Black patients, incorrectly producing risk assessments and, in effect, perpetuating health inequity [9]. A typical example is the case of the neural

network in Heidelberg, in 2016, which was trained with more than 100,000 photographs labelled either "malignant" or "benign," hence allowing it to identify melanomas from clinical images and even suggest diagnostic methods by itself; it was tested against 58 dermatologists, including 30 experts from 17 countries, and outperformed them, as 95% of melanomas and 82.5% of benign moles were correctly detected compared to the dermatologists' rates of 88.9% and 75.7%, respectively; when this was published in Annals of Oncology, the study was hailed as a bright future for artificial intelligence in medicine, calling for further refinement to bring such tools into the clinic; but buried in the limitations section was an important issue: over 95% of the training images showed white skin-a critical limitation to the broader generalizability of the model [10]. Another example is a Toronto-based startup that built an auditory test for Alzheimer's disease detection, however, the test kit worked for only fluent English speakers of a particular Canadian dialect [11].

This issue of bias within AI systems is not contained; rather, it is also an instance of broad, societal biases very often baked right into the base materials like training datasets; in most cases, an AI system is learning from past data, repeating patterns of discrimination that are or have been, present in most other areas of life, and this changes not only individual lives but also the landscape of social justice, equality, and even trust in technology at large [12]. Although, there have been past agitations about the integrity of preprogrammed decision-making systems [13], and with the thriving acceptance of machine learning in healthcare applications, these agitations have also stretched out to ML models' potential integrity bias [14]. It is important to understand the kind of biases involved in AI and their source, as that lets one know at what point one should intervene. The current paper aims to review two essential kinds of bias (data bias and algorithm bias) that cause discrimination in most AI applications.

1.1 Research Aims

This research will assess the role of diverse training datasets, analyse stakeholder perceptions, and develop strategies based on principles of fairness, accountability, and transparency for making the AI-driven health system nondiscriminatory and reliable.

1.2 Research Objectives

The objectives of the study include the following:

- Identify sources of bias in AI diagnostic tools.
- Assess bias influence on various under-represented population demographic groups.
- Gather information from stakeholders about AI fairness and transparency of strategies to mitigate the bias in AI applications in healthcare.
- Establish a framework that is anchored on FAT: Fairness, Accountability, and Transparency.

Despite the numerous existing literature concerning the research topic, very few works give any particular attention to underrepresented populations. Most of the studies regarding mitigating bias look at general bias mitigation strategies, mostly not specific to the needs of underrepresented populations, those that happen to be highly impacted by biased AI systems. Most of the current studies also do not actively involve several key stakeholders in the identification and addressing of those biases, such as patients and clinicians from underrepresented populations. This paper bridges the knowledge gaps by underlining the contribution of stakeholders from a diverse set of AI developers, clinicians, and patients. Furthermore, this research is important, as it deals with one of the crucial topics in the field of healthcare; assessing the notion of AI systems as nondiscriminatory and fair.

2.0 MATERIAL AND METHODS

2.1 Related Works

Empirical evidence of this has been present within the framework of AI diagnostics through a series of studies that have contributed to data diversity and algorithmic fairness, apart from approaches aimed at reducing biases to improve AI performance across underrepresented groups. Key works leading to the insights for this study are discussed below.

Bias in Diagnostic AI Tools and Contribution of Data Augmentation

Burlina *et al*., (2021) [15], examined data imbalance and domain generalization in AI-assisted diagnosis of DR. According to them, biased training datasets result in the biased diagnostic performances of AI, particularly concerning dark-skinned patients. This becomes possible by using the generated models for augmentation of their training dataset and has levelled the accuracy of detection across the demographic groups. Thus, it would mean that synthetic data augmentation could serve as an effective bias mitigation strategy.

On the other side, Raizman *et al*., (2024) [16], conducted a Kruskal-Wallis hypothesis test while quantifying some bias in AI training models along with several techniques that may decrease biases. Most of the key points that they derived show the way Color Jitter actually works with improvements in models leading to improvements in fairness across different skin types of humans.

Bias in AI-Driven Clinical Decision-Making

George *et al*., (2023) [17], studied algorithmic bias in the oncology prediction models. In this oncology dataset, over 28,000 patient cases were studied. Their overall model performance had an AUC of 0.8; however, racial, ethnic, and socio-economic disparities did crop up, and it was truly a wake-up call for the development of bias-aware AI.

The empirical analysis by Stanley *et al*., (2024) [18], of synthetic neuroimages of bias in AI models used for medical imaging was based on different techniques to create and inject artificial biases. According to the findings of the authors, reweighting considerably helps balance out bias within the convolutional neural networks (CNNs).

Bias Mitigation in AI-Generated Medical Advice

Sadat & Shakeri (2024) [19], restructured the concept of bias within AI medical suggestions by introducing the BiasMD and DiseaseMatcher datasets, which fine-tune language models in health. Subsequently, their derived EthiClinician model outperformed GPT-4 in both ethical reasoning and diagnostic accuracy, confirming that indeed, there could be an enhancement of fairness within the AI medical suggestion with tailored-made datasets.

Frameworks for Measuring and Reducing Bias

Gulamali *et al*., (2023) [20], discussed AEquity, a new metric with which to find racial bias and AI-driven diagnostic tools within the medical practice using multitask logistic regressions. They developed a task-agnostic method to find and mitigate biases in real-world medical datasets when performing diagnostic modelling on chest X-ray images.

Tay *et al*., (2022) [21], discussed the MLMB framework modelling various data-driven and algorithmic sources of bias. The authors discussed how embedding fairness metrics across several points of the AI model development lifecycle would provide transparency over and accountability of the decision-making processes.

Systematic Reviews of Bias in Clinical AI

A scoping review by Cary *et al*., (2023) [22], identified 109 studies related to bias in clinical algorithms and described common challenges in detecting, mitigating, and evaluating fairness in bias. They purported that even though most of these solutions were technical and operational, none was universally accepted as the best practice in reducing bias; thus, a need for context-specific AI fairness frameworks.

Ghai & Mueller (2022) [23], presented D-BIAS, an interactive visual interface for the detection and rectification of bias in AI models. This method combined human-in-the-loop auditing that could empower users to detect and correct unfair causal relationships in AI-driven decision-making.

Gender Bias in AI Diagnostics

Larrazabal *et al*., (2020) [24], mentioned the problem of the balance of gender in medical imaging datasets and showed how such unequal representation causes AI misdiagnosis in underrepresented gender groups. This, thus shows that the balance in datasets related to demographic groups will improve AI-driven medical diagnostics. This research will employ a quantitative approach, with a survey-based design where stakeholders give their views and ideas on AI bias, its mitigation, and how to develop an AI framework. Such a design is meant to allow structured data collection that can be quantitatively analysed. The current study will investigate the analyses of stakeholder input regarding the perception of AI bias and suggestions for improvement in fairness in AI diagnostic tools. In this regard, the data will be collected from the stakeholders through questionnaires, which will then be analysed quantitatively to outline patterns and trends in their responses.

2.1 Study Design, Sampling Method and Data Collection

This paper used a purposive sample for research purposes. In such a design, all subjects have had experiences in one way or another with what was directly connected with the research objective. Clinicians who had experience with AI diagnostic tools, patients from under-represented groups (such as black people whose diagnosis results were made through AI), and developers who built or implemented an AI-powered healthcare system.

It will, therefore, be well positioned to elicit from the respondents detailed insights into challenges and opportunities in biased AI systems, taking into consideration what is included and not included based on that fact. Some of the strategies used to recruit the participants entailed liaison with black community organisations, and professional network associations, which ensured that besides providing an assortment, the key stakeholder involvement was substantial. This gives depth to findings, enabling rich targeted perspectives to maintain efficiency; hence, the focus is on relevant participants only.

Data will be collected by questionnaires distributed to developers, clinicians, and patients. The questionnaires will comprise both closed-ended questions that shall draw a wide range of responses.

2.2 Data Analysis Method

The data analysis was accomplished with the Statistical Analysis Software (SAS) for:
Descriptive Statistics: The nature of the stakeholders and their responses will be described.
Frequency Analysis: How frequently certain opinions or perceptions about AI bias are mentioned?
Quantitative Analysis of Stakeholder Input: This will be based on the following themes in stakeholder input:
Perception of AI Bias*:* The perception of stakeholders concerning how bias in AI impacts healthcare outcomes.
Importance of Diverse Datasets*:* How stakeholders view the use of diverse training datasets in mitigating bias.
Proposed Mitigation Strategies: Strategies that stakeholders suggest for enhancing fairness in AI diagnostic tools.

2.3 THEORETICAL FRAMEWORK

2.3.1 Fat Framework

The theoretical framework that underpins this research study is informed by the founding principles of FAT, complemented by further additions regarding training set diversity and stakeholder involvement. That sort of combination may start to tease out some of the challenges these biases pose in terms of creating such systems, which would be fair and equitable and would be of benefit to a wide variety of diverse patient populations. The FAT framework is based on three key building blocks: fairness, accountability, and transparency. Key tenets are discussed with relevance, strengths, and limitations for this research.

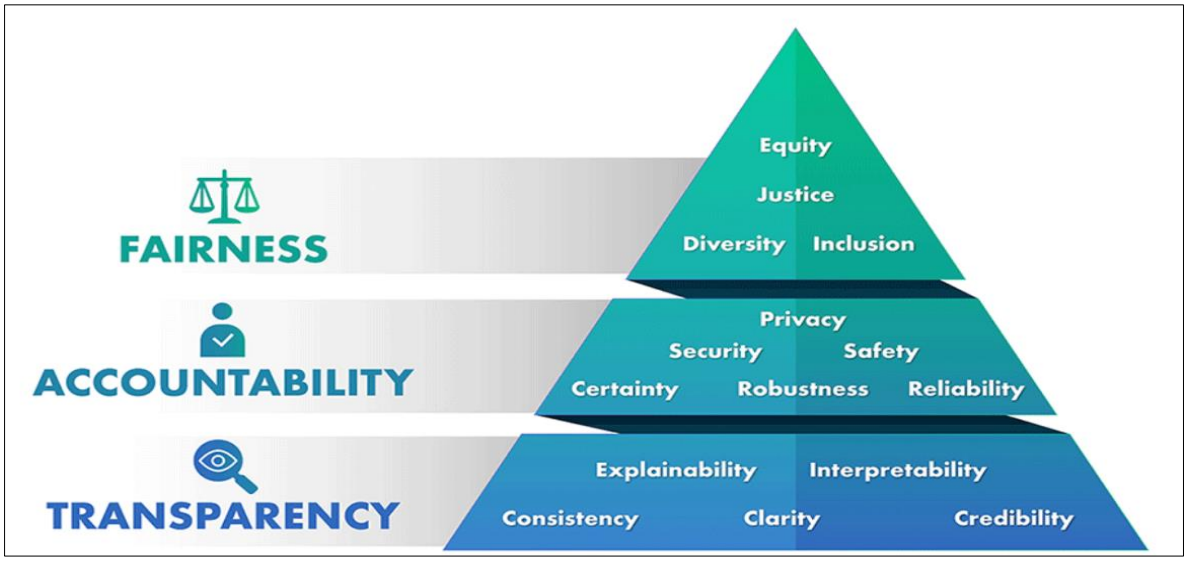

Figure 3: Theocharis. S (2023, February 8). *Understand AI Interpretability & Explainability - Towards AI*. Medium; Towards AI [25]

2.3.2 The First Pillar of FAT: Fairness- Fairness in AI Systems

Ethics of Fairness: Specifically, AI ethics have to make sure that decisions made by machines or computers are not discriminatorily unfair. As Giovanola and Tiribelli (2023) [26], noticed, fairness exceeds mere distributive equity or lack of discrimination; it relates more to the greater socio-relational plane that emanates from respect for persons in respect of the individuality of every single human person.

In other words, conceptualising fairness for AI systems is the equitable treatment of all or any one group for which the system operates within the FAT framework by reducing biases. What this means, as Whittaker *et al*., (2018) [27], and Selbst *et al*., (2019) [28], discuss, is a reduction in these biases. It matters because already disadvantaged groups could have disparities in diagnosis and treatment exacerbated within health domains. Deep fairness would, therefore, mean considering moral and ethical values, moving beyond bias but being fair to the creation and deployment of AI inclusively and respectfully.

2.3.3 The Second Pillar of FAT: Accountability- Accountability in AI Development

Accountability ensures that all stakeholders involved in the AI development lifecycle, such as the developers, the organisations, and others are held accountable for the effects of AI systems. This is based on the principle of well-defined roles, clear processes, and proactive measures in mitigating possible harm. Instead of after-the-fact blame casting, FAT accountability invites an approach that acknowledges and meets the responsibilities between developers and organisations regarding the development of accountable and responsible AI [29]. A culture of trust and integrity thus arises in which participants can flag some of the potential socio-impacts they may be creating through their work at an early stage.

Relevance:

Accountability can be the very basis on which AI diagnostic tools are simultaneously technically robust and socially responsible to address the bias that disproportionately hits the underrepresented population.

2.3.4 The Third Pillar of FAT: Transparency- Transparency by AI systems

Transparency will help build trust, understanding, and accountability among AI systems. According to Stoyanovich *et al*., (2018) [29], transparency allows AI systems to be more intelligible to the user and stakeholders through explainable decision-making procedures, algorithms, and data inputs. The transparent system allows scrutiny of the values output by AI, assessing their reliability, and identification of potential biases, thus providing users with a capacity for informed decision-making.

This would impact accountability and transparency that enables probing by stakeholders of such serious ethical and social consequences of these systems. For example, in health, it means that the clinicians and the patients can make sense of and thereby identify flaws or biases in how the AI diagnostic tools come up with recommendations on diagnosis.

Coupling the FAT framework with varied data and stakeholder engagement. This study would comprehensively put forth ways to handle bias in AI diagnostic tools. This would not only make AI systems fair, accountable, and transparent but also inclusive, ethical, and of benefit to all patient populations.

2.3.5 Application of the FAT Framework in this Research Study

The research framework provides the study not only with a conceptual but also an analytical approach wherein it gets executed and observed. At every stage in this study, right from the formulation stage to all the processes, it takes up the guiding principles of Fairness, Accountability, and Transparency to go forward with. It also serves as a guide to those stakeholders who are responsible and involved in the needed consideration of the provisions under the GDPR in their duty to protect individuals against undue effects [30]. This framework is included to ensure that, by its design, the study explores not only whether bias exists within AI-driven health but also whether, in such cases, it would be minimized and operationally applied. While fairness deals with researching how AI systems have different demographics in their effects and whether such systems act equitably across diverse populations, accountability drives the understanding of who is responsible for AI decisions, especially those that show which mechanisms may fail or work during an accountability stance in a life cycle development in AI. Finally, it is clear how the processes are communicated and understood by the users and if the stakeholders will be better informed with the capability to trust the system and use it [31].

The application of the FAT framework makes the methodology consonant with these core principles so that not only are biases identified but also recommendations about how they can be addressed within the processes of AI development and deployment. Thus, this will ensure that the findings are theoretically and practically applicable; hence, it will show how developers of AI and its stakeholders can encourage ethical AI in healthcare [31].

2.4 Data Synthesis and Analysis

A quantitative approach, using a survey-based instrument, has been undertaken in this study to accomplish the research objective, which is bias assessment and mitigation in AI diagnostic tools. In this way, the tool will receive measurable data from a wide variety of stakeholder groups, such as clinicians, patients, and AI developers, representing the underrepresented population. The current study operationalises perceptions of bias, fairness, and trust in the use of AI systems through Likert-scale questions; it builds, on a very firm systematic basis, the highlighting of challenges with actionable solutions. This format has allowed the survey tool to canvass many viewpoints on key issues such as the importance of diverse data, interdisciplinary collaboration, and transparency about methods regarding the handling of bias. The choices put forth were most apt for the trends across stakeholders, and, by putting stress on inclusivity and generalisability, goes to developing an equitable framework to help improve diagnostic tools on AI.

Statistical Analysis Software (SAS) was utilised in the analysis of the responses to the survey on perceived bias, fairness, and trust in AI diagnostic tools across demographic and professional categories. The contributions to the study had been 160 participants, hence a diverse demography.

3.0 RESULTS AND DISCUSSIONS

3.1 Descriptive Statistics

The data utilized in this research was collated from the relevant stakeholders (clinicians, AI developers and Patients), a total of 160 voluntary participants across certain ages, gender and professional demographics, as seen in Figure 1.1 below.

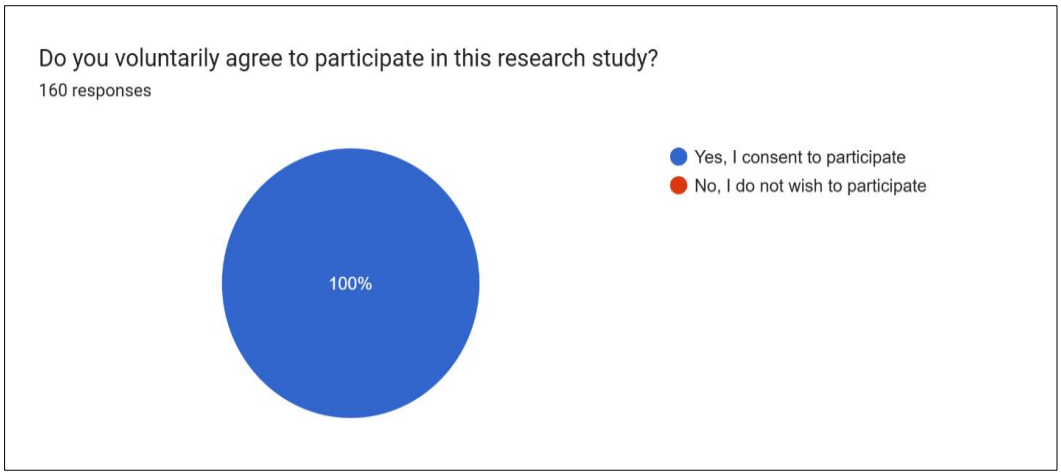

Figure 1.1: Total number of Participants in this survey

Demographic Variables

The largest proportion falls between 26–35 years old at 37.50%, followed by the 36-45 years old, accounting for 26.88%, with the older 46-55 years at 11.25%, 56-65 years at 5.63%, 66+ at 2.50%; and younger age groups 18-25 years at 16.25% having small representations as seen in Figure 1.2 below.

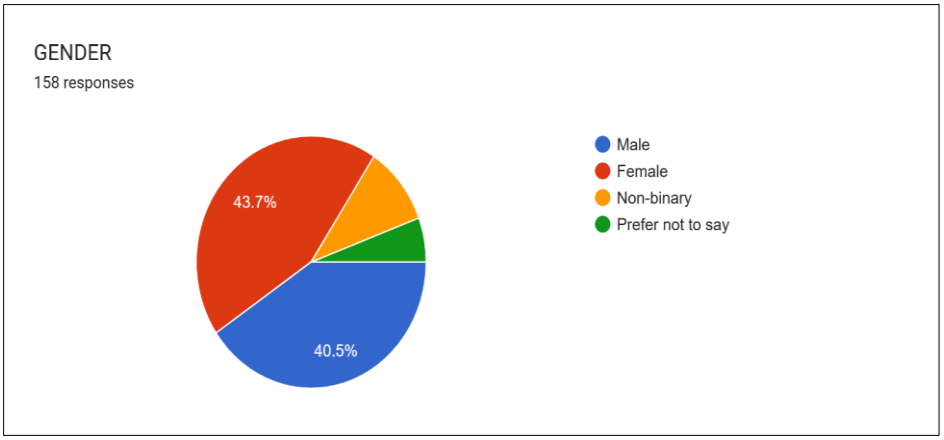

Figure 1.2: Total number of gender demographics in this survey

Concerning gender, there were 43.75% female, 40.63% male, 10.00% non-binary, while 5.63% preferred not to say, as seen in Figure 1.3 below.

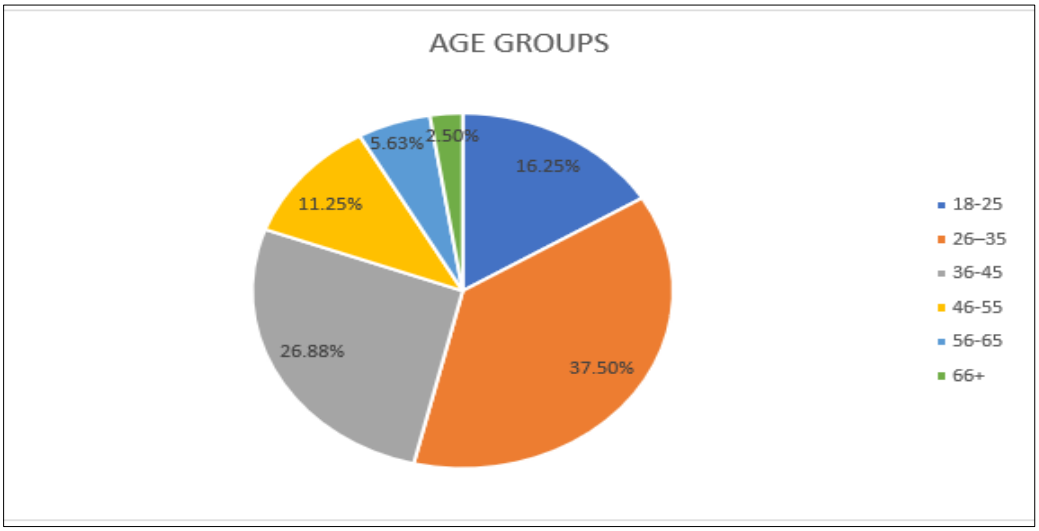

Figure 1.3: Total number of age groups of participants in this survey

The classification of participants in this survey includes healthcare patients 46.88%, clinicians 28.13%, and 25.00% AI developers, among others were also included, as seen in Figure 1.4 below.

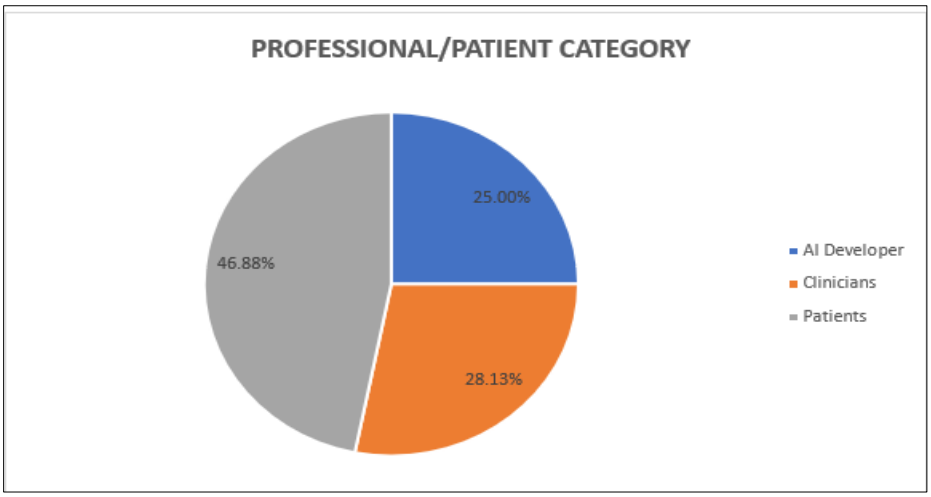

Figure 1.4: Total number of stakeholders demographics in this survey

Likert-Scale Responses

Ranges of response are between 1 to 5 (with Strongly Disagree as 1; Disagree as 2; Neutral as 3; Agree as 4; and Strongly Agree as 5).

Research Objective 1

Table 1: Bias in training datasets is the primary cause of fairness issues in AI diagnostic tools.

	Frequency	Per cent	Cumulative Frequency	Cumulative Percent
1	2	4.76	2	4.76
2	6	14.29	8	19.05
3	12	28.57	20	47.62
4	20	47.62	40	95.24
5	2	4.76	42	100.00

Table 1.1

Statement	Mean	Standard deviation
Bias in training datasets is the primary cause of fairness issues in AI diagnostic tools.	3.33333	0.95424

A higher percentage (47.62%) of the stakeholder responses agreed to the question: *Bias in training datasets is the primary cause of fairness issues in AI diagnostic tools* as seen in Tables 1. Also, the mean being 3.33 and a lower standard deviation (0.95) as seen in Table 1.1, support the frequency data that most of the responses relate biased training data as an influential factor leading to problems in fairness-related aspects, reflecting that the stakeholders agree with the statement.

Table 2: Bias in AI diagnostic tools poses a significant challenge to clinical decision-making.

	Frequency	Per cent	Cumulative Frequency	Cumulative Percent
1	5	11.11	5	11.11
2	4	8.89	9	20.00
3	19	42.22	28	62.22
4	14	31.11	42	93.33
5	3	6.67	45	100.00

Table 2.1

Statement	Mean	Standard deviation
Bias in AI diagnostic tools poses a significant challenge to clinical decision-making.	3.13333	1.05744

Furthermore, 33.11% agreed to the question: *Bias in AI diagnostic tools poses a significant challenge to clinical decision-making, suggesting that bias creates big challenges in AI diagnostic tools* as seen in Tables 2 below. While, supporting the frequency data, the mean of 3.13 as seen in Table 2.1 presents a higher degree of stakeholder agreement

that one of the most important challenges facing clinical decision-making is bias, and the standard deviation, which stands at 1.06, being somewhat low presents a shared view that AI bias indeed is one of the main barriers.

Research Objective 2

A higher percentage (40%) of the stakeholder responses agreed to the question: AI diagnostic tools are generally reliable across diverse patient populations. However, with an average of 3.04, which is close to neutral, the stakeholders do not have much belief in the ability of AI diagnostic tools to perform uniformly across different demographics. Further, a good dispersion marked by a standard deviation of 1.13 shows that not all are confident in the consistency of AI performance across diverse populations as seen in Tables 3 and 3.1.

Table 3: AI diagnostic tools are generally reliable across diverse patient populations.

	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	5	11.11	5	11.11
2	10	22.22	15	33.33
3	10	22.22	25	55.56
4	18	40.00	43	95.56
5	2	4.44	45	100.00

Table 3.1

Statement	Mean	Standard deviation
AI diagnostic tools are generally reliable across diverse patient populations.	3.13333	1.12726

Furthermore, 48.15% agreed to the question: *Bias in AI diagnostic tools affects the quality of healthcare I receive*, highlight the disparities in perceived fairness across different demographic groups as seen in **Table 4** below. The high mean of 3.5 supports that stakeholders generally agree that Bias in AI diagnostic tools affects the quality of healthcare they receive, while the standard deviation of 0.9, being low suggests a good agreement among stakeholders that bias in AI diagnostic tools affect the quality of healthcare as reflected in Table 4.1.

Table 4: Bias in AI diagnostic tools affects the quality of healthcare I receive

	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	4	4.94	4	4.94
2	8	9.88	12	14.81
3	21	25.93	33	40.74
4	39	48.15	72	88.89
5	9	11.11	81	100.00

Table 4.1

Statement	Mean	Standard deviation
Bias in AI diagnostic tools affects the quality of healthcare I receive.	3.50617	0.98898

Research Objective 3

In assessing stakeholder perception, diverse perception on fairness in AI diagnostic tools and dataset in underrepresented populations from all the stakeholders involved in this survey were captured in Table 5, 6 and 7 below. As reflected in Table 5, 31.11% of the stakeholder responses were neutral to the question: *Do you agree that biases in AI diagnostic tools impact patient outcomes negatively.* Supporting this results as seen in Table 5.1, the mean relatively in the middle, 2.78 explains that the stakeholder respondents somewhat agreed to AI bias having negative impacts on the patient outcome though the standard deviation being 1.18 reveals noticeable variation among respondents, thus some strongly agreed and neutrality and disagreements were present.

Table 5: Do you agree that biases in AI diagnostic tools impact patient outcomes negatively?

	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	7	15.56	7	15.56
2	12	26.67	19	42.22
3	14	31.11	33	73.33
4	8	17.78	41	91.11
5	4	8.89	45	100.00

Table 5.1

Statement	Mean	Standard deviation
Biases in AI diagnostic tools impact patient outcomes negatively.	2.77778	1.18492

Furthermore, 38.6% of the stakeholder responses were neutral to the question; *I trust AI diagnostic tools to provide fair and accurate assessments for my patients* as seen in Table 6 below, with the mean being 3.18 indicates rather moderate levels of trust in AI regarding fairness and accuracy, and an SD of 0.97, showing there is slightly higher uniformity of response, and the trust in fairness and accuracy in diagnosis by means of AI is moderate as seen in Table 6.1 below**.**

Table 6: I trust AI diagnostic tools to provide fair and accurate assessments for my patients

	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	2	4.55	2	4.55
2	8	18.18	10	22.73
3	17	38.64	27	61.36
4	14	31.82	41	93.18
5	3	6.82	44	100.00

Table 6.1

Statement	Mean	Standard deviation
I trust AI diagnostic tools to provide fair and accurate assessments for my patients.	3.18182	0.97104

Moreover, 59.52% of the stakeholder agreed to the question: *Stakeholder feedback is crucial for improving fairness in AI diagnostic tools* as seen in Table 7. The high mean score of 3.74 infers that input from the stakeholders is required in refining the fairness of AI, while the moderate standard deviation of 0.94 as seen in Table 7.1 shows variance in stakeholders opinion; however, agreements do exist.

Table 7: Stakeholder feedback is crucial for improving fairness in AI diagnostic tools.

	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	2	4.76	2	4.76
2	2	4.76	4	9.52
3	7	16.67	11	26.19
4	25	59.52	36	85.71
5	6	14.29	42	100.00

Table 7.1

Statement	Mean	Standard deviation
Stakeholder feedback is crucial for improving fairness in AI diagnostic tools.	3.73810	0.93859

Research Objective 4

In understanding the ways bias could be reduced, 53.33% of the respondents agreed that diversity of data enhances fairness in AI diagnostic tools as seen in Table 8, supporting this result is the mean at 3.64, which is a relatively high number; thus signifies very strong agreement that dataset diversity indeed plays an important role in fairness, while the standard deviation is 1.05, and thus this suggests that the agreement by all the stakeholders on the inclusivity of AI training datasets as seen in Table 8.1.

Table 8: Diverse datasets are crucial for improving the fairness of AI diagnostic tools.

	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	3	6.67	3	6.67
2	3	6.67	6	13.33
3	8	17.78	14	31.11
4	24	53.33	38	84.44
5	7	15.56	45	100.00

Table 8.1

Statement	Mean	Standard deviation
Diverse datasets are crucial for improving the fairness of AI diagnostic tools.	3.64444	1.04785

Table 9 below shows that 76.88%, also agreed that regular audits of the algorithms help in enhancing the fairness of AI systems, the high mean of 3.8 also indicates that stakeholders generally agree that regular algorithmic audits are

crucial for identifying and addressing potential biases in AI dataset, while the standard deviation of 0.7 as shown in Table 9.1 suggests a strong agreement among the respondents.

Table 9: Regular algorithmic audits are crucial for identifying and addressing potential biases

	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	4	2.50	4	2.50
2	7	4.38	11	6.88
3	18	11.25	29	18.13
4	123	76.88	152	95.00
5	8	5.00	160	100.00

Table 9.1

Statement	Mean	Standard deviation
Regular algorithmic audits are crucial for identifying and addressing potential biases.	3.77500	0.71770

In Table 10 below, 60% agreed to the suggestion of interdisciplinary collaboration helping in enhancing the fairness of AI systems, supporting this results is an average mean of 3.55 which indicates that collaboration across disciplines is a good method of improving equity, while a standard deviation of 1.04 as reflected in Tables 10.1 shows tremendous diverse opinions while the general agreement remains.

Table 10: Interdisciplinary collaboration significantly improves AI fairness.

	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	6	3.75	6	3.75
2	11	6.88	17	10.63
3	35	21.88	52	32.50
4	96	60.00	148	92.50
5	12	7.50	160	100.00

Table 10.1

Statement	Mean	Standard deviation
Interdisciplinary collaboration significantly improves AI fairness.	3.54762	1.04069

Research Objective 5

In Table 11 below, a high percentage of respondents (51.25%) agreed that transparent AI development processes help mitigate bias effectively in AI diagnostic tools; the mean of 3.66 supports the frequency results that transparency is one of those huge solutions that will reduce bias, while the standard deviation of 1.00 as seen in Table 11.1 infers strong agreement.

Table 11: Transparent AI development processes help mitigate bias effectively.

	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	7	4.38	7	4.38
2	15	9.38	22	13.75
3	30	18.75	52	32.50
4	82	51.25	134	83.75
5	26	16.25	160	100.00

Table 11.1

Statement	Mean	Standard deviation
Transparent AI development processes help mitigate bias effectively.	3.65625	1.00343

Also, in Table 12 below, 42.22% of the stakeholder responses were neutral to the question: *current fairness frameworks adequately address the issue of bias in AI diagnostic tools*. These results reflect the sensitivity rising in this issue of offsetting biases in AI systems. In support of the frequency data results, 3.24, being the mean as seen in Table 12.1 explains that while few of the respondents may believe that the fairness frameworks are good, many are sceptical as to how much effort is being put in. The standard deviation also being 1.11; explains small deviations in perception with regards to how effective the current fairness frameworks really are, or if these even exist.

Table 12: Current fairness frameworks adequately address the issue of bias in AI diagnostic tools.

	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	6	13.33	6	13.33
2	2	4.44	8	17.78
3	15	33.33	23	51.11
4	19	42.22	42	93.33
5	3	6.67	45	100.00

Table 12.1

Statement	Mean	Standard deviation
Current fairness frameworks adequately address the issue of bias in AI diagnostic tools.	3.24444	1.11101

Conclusively, from the statistical analysis in the Tables above; there were high levels of agreement on bias issues, as most stakeholders agreed that AI bias hurts patient outcomes, showing that AI fairness is of interest to underrepresented populations. Also, there is a need for diverse training datasets; as many responded that AI datasets need to have some representative data in the training sets. The different stakeholders surveyed have different levels of trust in using AI tool systems while there was wide agreement among the respondents that regular audits are important with the implementation of AI models. However, in the analysis of gaps in existing fairness frameworks; the stakeholder's respondent showed that reflected incompleteness regarding the available fairness frameworks about AI diagnostics.

3.2 Thematic Areas

- **Perception of Bias in AI Diagnostic Tools Outcomes:** An overwhelming number of respondents reported that there was perceived bias in the findings from the AI diagnostic system. Algorithmic discrimination against health professionals and patients was among the overwhelming challenges brought forth while making clinical decisions.

Interpretation: This confirms evidence from research studies that prejudicial training datasets lead to prejudicial diagnostic performance, and bias in AI diagnostic tools encourages health care disparities disproportionately between ethnic minorities and gender groups [24].

- **Impact of Bias on Under-represented Populations:** All participants from underrepresented demographics in this survey tend to have lesser trust in the diagnosis by AI diagnostic tools. For example, the female gender reports that AI produces more errors than others. Generally, all participants felt that the data from which training was done was so poorly representative. This result indicates the urgent need to advance diverse datasets and interventions regarding the assurance of fairness.
- **Stakeholder Trust in AI Fairness:** Variations existed between clinicians and developers in trusting the diagnosis by AI, while a portion of the respondents felt that mechanisms of transparency and accountability are not sufficient as far as AI systems are concerned.

Interpretation: Another significant moderator in the building of stakeholder trust is transparency. According to Stoyanovich *et al*., (2018) [29], AI systems have to be transparent in their decision-making for wide acceptance.

- **Bias Mitigation Strategies:** Most of the responses point toward diverse training data sets as an important strategy for the reduction of bias. Overall, both have been fundamentally accepted as what has been adopted and are widely used in ensuring fairness across AI systems. Other participants indeed suggested an interdisciplinary approach in upgrading the framework in terms of improvement toward being fair. Interpretation: Based on Raizman *et al*., (2024) [16], bias mitigation strategies involve dataset augmentation or reweighting, which results in the reduction of disparities in diagnosis.

Conclusively, this thematic analysis outlines a synthesis of findings concerning the challenges and stakeholders' concerns about possible solutions concerning bias problems in AI diagnostic tools. These were supported by the calls for data, fairness metrics, and transparency and regulatory policies that can help build an equitable AI-driven health system.

4.0 CONCLUSION

This paper addresses one of the critical challenges that arise from bias in AI diagnostic systems. The author investigated discrimination within healthcare technologies and considered how using a diverse and representative training dataset can be used to help improve health equity for underrepresented populations. Using a quantitative approach, it collected questionnaires from the key stakeholders in healthcare and AI development through structured questionnaires that captured perceptions of AI bias as well as strategies for the mitigation of AI bias. This study bridges the knowledge gaps of sources and impacts of bias by gathering insights from stakeholders and frameworks to mitigate algorithmic discrimination. This translates to hypothesising that training data diversity reduces bias in medical applications of AI

systems, consequently producing fewer undesirable health outcomes among underrepresented populations while providing implementable recommendations on the development of such tools, which can consider equity in mind.

These studies on bias in AI systems, particularly on diagnostics in health care, will therefore carry critical insight into demanding an altogether more holistic response. Put together, these studies bring to the fore, the highly debilitating effects that algorithmic bias serves only to heighten existing disparities in health care. Another important fact is that AI systems themselves are not neutral, and may reproduce discriminative patterns of society, especially when their training was conducted with data that did not represent the population.

This is important to show, based on the previous discussion, that the discussion of AI diagnostic tool bias and its reduction with the intent of making health outcomes nondiscriminatory is direly needed for all, but most especially for underrepresented populations. With this, it will be able to elaborate on how biases in the training dataset and algorithms impact the accuracy and reliability of the AI systems themselves in ways that could provide a basis for disparities in healthcare delivery. This work epitomizes, concisely, how the stakeholders view ranges, from clinicians to patients, and even the very developers of AI themselves showing key strategies for improving fairness and transparency in diagnostic AI through data diversity, periodic audits of algorithms, and interdisciplinary collaboration.

Thus, these insights have set up robust frameworks aimed at depleting biases and building trust in systems working effectively and equitably toward patient populations.

4.1 Policy Implications

These research findings could have important implications for policy in the development and deployment of AI diagnostic tools, in that these companies are under pressure from policymakers to include varied and representative datasets while training and testing AI models to avoid biased performance that would differentially affect underrepresented populations.

First, this study highly advocates multi-dimensional systematic interventions. In this regard, policymakers and health institutions must concentrate on representative training data sets that are diverse and incorporate nuanced variation across demographic groups. It has nothing to do with the numbers themselves but how these normally underrepresented populations get meaningfully included in the development of these models of AI.

Second, AI developers are to set up periodic algorithmic audits to uncover possible biases in datasets or diagnostic tools.

Third, there is a need to develop standards for the transparency and explainability of AI systems that can help build trust in various stakeholders, including patients, clinicians, and developers of AI themselves, to understand and make judgments about AI-driven decisions. This calls for collaboration across regulatory bodies, healthcare institutions, and AI developers in developing policies that emphasize ethical AI development with accountability in consideration.

5.0 LIMITATIONS

The present study has gone a long way in deepening the understanding of bias in AI diagnostic tools and their mechanisms for the framework of fairness evaluation. There are, however, a certain number of limitations that have to be taken into consideration. These limitations also present avenues through which potential improvements might be considered for future studies.

Sample Size and Generalisability:

Although heterogeneous, the sample is not fully representative, in particular of all health stakeholders in low-resource settings where the deployment of AI is still in its relative infancy. Future studies should involve larger participant groups that are more geographically diversified to allow increased generalisability of findings.

Self-Reported Data Bias:

Because it is a survey study, dependence on self-reporting is necessarily prone to biases of response through which participants tend to see over or underestimate events of experience of bias from the AI. Further studies can adopt an observational or experimental approach to determine whether AIs are biased or not biased. These analyses are objective.

Limited AI Domain Coverage:

The current study centres on diagnostic AI, while other applications of the use of AI in healthcare are not part of this research work, such as predictive analytics, robotic surgeries, and administrative automation. Bias across a more diverse range of AI applications, investigated in a broader range of healthcare settings can be incorporated in future research.

Lack of Direct Testing of AI Models:

This research does not report any empirical testing of AI models for bias regarding real-world diagnostic performance. Future studies should be done with algorithmic audits, testing fairness with diverse real-world medical datasets.

5.3 Recommendations

These recommendations are based on the following challenges identified during the investigation undertaken in this study.

First, there is also a need for regular audits by institutions in which an AI system is deployed concerning algorithm performance. In other words, the development of such AI systems for real-world healthcare needs will involve the cooperation of its developers, clinicians, and patients to work with one another. This would include applying tight data augmentation techniques for syntheses of diversity within the datasets and representation particularly the marginalized frameworks for the assessment of bias, such frameworks would afford uniform measurements regarding the fairness of an algorithm toward various subgroups oversight through a mechanism that includes, amongst others clinicians, AI developers, ethicists, and patients' representatives to continuously review and improve on diagnostic AI tools.

Secondly, I recommend standardization of all AI models; where each model is to undergo processes for diversity and inclusion; so that they can be ingrained deep within all races, ethnic groups, genders, and socio-economic backgrounds.

Furthermore, for transparency measures, the documentation of AI systems should be in a transparent manner with feature explainability for illustrating to stakeholders their basis upon AI-driven decisions.

Moreover, clinicians should undergo regular training on the use of AI systems and their biases, which will be very important for healthcare providers as they begin the use of these diagnostic tools effectively and ethically. Also, regular reports on how AI diagnostic systems are working with different populations should be given by clinicians to help AI developers monitor, improve, and extend AI diagnostic tools.

Finally, based on this research findings, there is a need to stimulate novel, nondiscriminatory, and reliable diagnostic AI; as this will ensure technology as a driver of equitable access to healthcare and improved health outcomes for all populations.

5.4 Future Research Directions

The following improvements are recommended for future research

Extension of Methodologies:

A machine learning model should be incorporated for the evaluation and testing of AI diagnostic tools on various balanced versus imbalanced datasets, measuring changes in fairness.

The use of mixed-method approaches to combine insightful qualitative research findings, such as clinician interviews, with quantification of the performance of AIs.

Longitudinal Studies:

Longitudinal studies tracking AI over time in their fairness, such as when available datasets themselves become more heterogeneous, and also when mechanisms to ensure fairness were developed and then implemented can be conducted.

Policy and Regulatory Analysis:

Future work should do more to explain better how AI regulations and ethics broadly affect the fairness of deployed AI in the real-world setting such as policies that work to reduce bias for different countries.

Conclusively, expanding these research methodologies in future studies will go a long way towards the enhancement of fairness, transparency, and accountability in AI healthcare applications.

Acknowledgements

Conflicts of Interest

Ethical Approval

The informed consent of the participants in this study had been obtained in advance, and thus the form of consent appeared as an introduction to the online questionnaire explaining the following:

- The purpose of the study,
- Their voluntary participation,

- Their right to withdraw at any time,
- The confidentiality of their responses.

By continuing with the questionnaire, participants gave their explicit consent to participate. No personal data were collected that could identify participants, so privacy and anonymity were guaranteed.

REFERENCES

1. Chinta, S. V., Wang, Z., Zhang, X., Viet, T. D., Kashif, A., Smith, M. A., & Zhang, W. (2024). Ai-driven healthcare: A survey on ensuring fairness and mitigating bias. *arXiv preprint arXiv:2407.19655*.
2. Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial intelligence and human trust in healthcare: focus on clinicians. *Journal of medical Internet research*, *22*(6), e15154.
3. Balasubramanian, S., Devarajan, H. R., Raparthi, M., Dodda, S. B., Maruthi, S., & Adnyana, I. M. D. M. (2023). Ethical Considerations in AI-assisted Decision-Making for End-Of-Life Care in Healthcare. *Power System Technology*, *47*(4), 167-182.
4. James, T. A. (2024). Confronting the mirror: reflecting on our biases through AI in health care [Internet]. *Harvard.edu*. Sep 24 [cited 2024 Dec 01]. Available from: https://postgraduateeducation.hms.harvard.edu/trends-medicine/confronting-mirror-reflecting-our-biases-through-ai-health-care
5. Hanna, M., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., ... & Rashidi, H. (2024). Ethical and Bias considerations in artificial intelligence (AI)/machine learning. *Modern Pathology*, 100686.
6. Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns*, *2*(10).
7. Cross, J. L., Choma, M. A., & Onofrey, J. A. (2024). Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health*, *3*(11), e0000651.
8. Mittermaier, M., Raza, M. M., & Kvedar, J. C. (2023). Bias in AI-based models for medical applications: challenges and mitigation strategies. *NPJ Digital Medicine*, *6*(1), 113.
9. Stetler, C. (2024). AI algorithms used in healthcare can perpetuate bias [Internet]. Rutgers.edu; Available from: https://www.newark.rutgers.edu/news/ai-algorithms-used-healthcare-can-perpetuate-bias
10. Dutchen, S. (2019). The importance of nuance [Internet]. *Artificial Intelligence Issue*. [cited 2024 Dec 01]. Available from: https://magazine.hms.harvard.edu/articles/importance-nuance
11. Gershgorn, D. (2018). If AI is going to be the world's doctor, it needs better textbooks [Internet]. *QUARTZ Daily Brief*. [cited 2024 Dec 01]. Available from: https://qz.com/1367177/if-ai-is-going-to-be-the-worlds-doctor-it-needs-better-textbooks
12. Lewis, J., Lopez, M., & Ramachandran, H. (2024). AI and bias: addressing discrimination in machine learning algorithms. *AlgoVista J AI Comput Sci*, *1*(2).
13. Hutchinson, B., & Mitchell, M. (2019, January). 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 49-58).
14. Chen, I. Y., Szolovits, P., & Ghassemi, M. (2019). Can AI help reduce disparities in general medical and mental health care?. *AMA journal of ethics*, *21*(2), 167-179.
15. Burlina, P., Joshi, N., Paul, W., Pacheco, K. D., & Bressler, N. M. (2021). Addressing artificial intelligence bias in retinal diagnostics. *Translational Vision Science & Technology*, *10*(2), 13-13.
16. Raizman, E., Peng, Y., & Deng, A. (2024, August). Towards Equitable Diagnosis: Bias Evaluation and Mitigation in Skin Cancer Classification. In *2024 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* (pp. 1-9). IEEE.
17. George, R. D., Ellis, B. H., Sidey-Gibbons, C. J., & Swisher, C. (2023). Protecting against algorithmic bias of AI-based clinical decision making tools in oncology. *Cancer Research*, *83*(7_Supplement), 1970-1970.
18. Stanley, E. A., Souza, R., Winder, A. J., Gulve, V., Amador, K., Wilms, M., & Forkert, N. D. (2024). Towards objective and systematic evaluation of bias in artificial intelligence for medical imaging. *Journal of the American Medical Informatics Association*, *31*(11), 2613-2621.
19. Sadat Zahraei, P., & Shakeri, Z. (2024). Detecting Bias and Enhancing Diagnostic Accuracy in Large Language Models for Healthcare. *arXiv e-prints*, arXiv-2410.
20. Gulamali, F. F., Sawant, A. S., Liharska, L., Horowitz, C. R., Chan, L., Kovatch, P. H., ... & Nadkarni, G. N. (2023). An AI-Guided Data Centric Strategy to Detect and Mitigate Biases in Healthcare Datasets. *arXiv preprint arXiv:2311.03425*.
21. Tay, L., Woo, S. E., Hickman, L., Booth, B. M., & D'Mello, S. (2022). A conceptual framework for investigating and mitigating machine-learning measurement bias (MLMB) in psychological assessment. *Advances in Methods and Practices in Psychological Science*, *5*(1), 25152459211061337.
22. Cary Jr, M. P., Zink, A., Wei, S., Olson, A., Yan, M., Senior, R., ... & Pencina, M. J. (2023). Mitigating racial and ethnic bias and advancing health equity in clinical algorithms: a scoping review: scoping review examines racial and ethnic bias in clinical algorithms. *Health Affairs*, *42*(10), 1359-1368.

23. Ghai, B., & Mueller, K. (2022). D-bias: A causality-based human-in-the-loop system for tackling algorithmic bias. *IEEE Transactions on Visualization and Computer Graphics*, *29*(1), 473-482.
24. Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., & Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, *117*(23), 12592-12594.
25. Theocharis, S. (2023). Understand AI interpretability & explainability - Towards AI [Internet]. Medium; Feb 08. Available from: https://pub.towardsai.net/understand-ai-interpretability-explainability-3718797777ec
26. Giovanola, B., & Tiribelli, S. (2023). Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. *AI & society*, *38*(2), 549-563.
27. Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., & Schwartz, O. (2018). AI Now 2018 Report [Internet]. New York: AI Now Institute; [cited 2024 Dec 01]. Available from: https://ainowinstitute.org/publication/ai-now-2018-report-2
28. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019, January). Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 59-68).
29. Stoyanovich, J., Howe, B., Jagadish, H. V., & Miklau, G. (2018). Panel: a debate on data and algorithmic ethics. *Proceedings of the VLDB Endowment*, *11*(12), 2165-2167.
30. Blacklaws, C. (2018). Algorithms: transparency and accountability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2128), 20170351.
31. Hoffmann, A. L., Roberts, S. T., Wolf, C. T., & Wood, S. (2018). Beyond fairness, accountability, and transparency in the ethics of algorithms: Contributions and perspectives from LIS. *Proceedings of the Association for Information Science and Technology*, *55*(1), 694-696.