American Journal of Humanities and Social Sciences Research (AJHSSR)
e-ISSN: 2378-703X
Volume-09, Issue-03, pp-176-186
www.ajhssr.com
Research Paper Open Access

AI FOR EQUITY: DEVELOPING ADAPTIVE BIAS DETECTION FRAMEWORKS FOR HEALTHCARE ALGORITHMS

Adeola Bamiro
University of the Cumberlands. Department of Information Technology

ABSTRACT
Objective: Healthcare AI models possess substantial abilities to enhance the quality of patient outcomes while identifying medical diagnoses and treatment methodologies. The existing biases which exist within AI models produce unequal medical outcomes across different population groups who face discrimination in their care. The research focuses on creating adaptable methods and uniform assessment criteria for healthcare AI bias detection methods and their bias reduction strategies.
Materials and Methods: A comprehensive study of current AI bias detection methods occurred while examining crucial instances of healthcare algorithms showing bias that resulted in unequal patient results. A bias detection framework involving continuous monitoring and both explainable AI (XAI) and regulatory compliance exist to examine the system at multiple levels. The system creates a new method to measure both the magnitude of bias and its effects.
Results: The initial experimental evidence shows that classic bias correction methods do not suit changes in actual data distributions. The introduced framework achieves higher accuracy in bias detection according to preliminary testing since it decreases disparate impact scores of chosen AI models by 35%.
Discussion: These results show that we need adaptive methods for bias reduction which should match the development of AI models and datasets. This paper analyzes predictive healthcare model bias together with associated ethical matters and regulatory elements that influence fair AI implementation.
Conclusion: The establishment of adaptable systems that detect bias promotes the development of responsible AI solutions in healthcare. Standardized metrics for bias assessment will establish confidence between users and healthcare providers while minimizing inequalities in AI-driven healthcare services.

I. INTRODUCTION

1.0 Background of the Study

Artificial Intelligence has brought revolutionary changes to healthcare through its diagnostic systems and management programs for patients and clinical decision processes. The models demonstrate biased outputs because they use flawed data distribution alongside an insufficient representation of minority groups and mathematical system errors. Research shows that AI-based disease prediction tools fail to provide precise evaluations to specific racial and socioeconomic populations because these groups receive inaccurate diagnosis results, which create unfavorable medical consequences. [1]AI growth requires protective systems for both detecting and fixing biases since this would stop the healthcare system from worsening social disparities between demographic groups. Healthcare AI faces regulatory challenges because standard metrics combined with dynamic frameworks are not yet available for detecting emerging biases according to the EU AI Act and U.S. FDA guidelines. [2,3] The proposed research aims to establish such a framework for bias detection.

1.1 Statement of the Problem

Being biased within healthcare AI systems proves to be a major challenge that violates both fairness and equitable standards in medical practice. The diagnostic errors from non-diverse training datasets, together with underestimated treatment needs, result in both medical and healthcare equity issues. The current methods for detecting bias incorporate static fairness regulations, which demonstrate incompetence in following data pattern modifications. Currently there exists no common agreement on measuring or tracking biases in a method consistent with effective correction. This study tackles these weaknesses through the creation of an adaptive standard method which releases AI-powered healthcare systems from biases.

1.2 Objectives of the Study

The primary objective of this study is to design a dynamic bias detection and mitigation framework for healthcare AI applications. Specifically, the study aims to:

- Identify the key sources of bias in AI-driven healthcare algorithms.
- Develop a standardized metric system that can be used to quantify the impact and severity of bias.
- Design an adaptive framework that monitors and corrects biases continuously.
- Determines how effective the proposed framework is in reducing disparities in healthcare AI models.
- Recommend best practices for equitable AI deployment in medical settings.

1.3 Research Questions:

- What are the primary causes of bias in healthcare AI models?
- How can the severity and impact of bias be systematically quantified?
- What components should an effective bias detection framework include?
- How does an adaptive bias mitigation approach improve fairness in AI-driven healthcare applications?
- What best practices can be implemented to ensure equity in AI healthcare deployment?

1.4 Relevant Hypothesis:

- **H1:** The proposed adaptive framework will enhance bias detection accuracy in healthcare AI models in a significant way.
- **H2:** When implemented, standardized bias quantification metrics will help to improve transparency and fairness in AI-driven healthcare solutions.
- **H3:** Continuous monitoring and real-time bias mitigation will reduce healthcare disparities across different demographic groups.

1.5 Significance of the Study

This study is significant for multiple stakeholders in healthcare, technology, and policy:

- Amongst healthcare practitioners, it will help to ensure AI-driven diagnostics and treatments are equitable.
- It will provide AI developers with a structured framework for bias assessment and mitigation.
- It offers guidelines for fair AI governance in medical applications.
- Enhancing patient's trust and reliability in AI-powered healthcare decisions.

By addressing bias proactively, this research aims to foster more ethical, equitable, and effective AI deployment in healthcare.

1.6 Scope of the Study

The analysis concentrates on confirming and resolving biases in medical diagnosis AI systems in combination with predictive analytics and patient care systems. The practice includes three components of bias generation that stem from training datasets together with model architectures and deployment environments. The research analysis excludes the examination of bias found in medical technologies that do not use AI and the systems through which healthcare administration occurs.

1.7 Definition of Terms

- **Bias in AI:** Systematic deviation in AI model predictions that leads to unfair outcomes for specific demographic groups.
- **Adaptive Bias Detection:** A dynamic approach to identifying and mitigating bias that evolves alongside changes in data patterns and model behavior.
- **Explainable AI (XAI):** AI models designed to provide transparency into their decision-making processes.
- **Fairness Metrics:** Quantitative measures used to evaluate equity in AI predictions, such as disparate impact ratio and equalized odds.
- **Healthcare AI:** Machine learning and AI-based systems used for diagnosis, treatment recommendations, and medical research.

II. LITERATURE REVIEW

2.1 Preamble

The modern age witnesses Artificial Intelligence (AI) as a groundbreaking technology that reformulates numerous industries such as finance, transportation, security, and education but shows its greatest impact within healthcare. Manufacturers of healthcare technology use artificial intelligence to improve healthcare delivery through diagnosis systems and patient tracking mechanisms, medication recommendations, and new medication identification. The combination of machine learning algorithms alongside natural language processing and predictive analytics enables advanced healthcare systems to carry out complex clinical choices, enhance diagnosis, forecast patient responses, and customize medical treatments at a superior speed compared to traditional protocols.[4-6] These developments lead to accelerated workflows in addition to leading to better

healthcare results together with decreased healthcare expenses as well as better operational results in medical practice. Strong healthcare system dependence on AI integration has generated essential ethical matters plus social challenges and technical barriers among which algorithmic bias emerges as a foremost concern. Systemic errors that lead to unfair and unfavorable outcomes within Artificial Intelligence decision systems constitute algorithmic bias. These biases manifest from training data, which exposes historical and societal prejudices and systemic inequalities within modern society. [7-8] The penetration of biases within healthcare algorithms increases health inequalities by intensifying current disparities, which mainly affect underprivileged and disadvantaged communities. AI systems that work with bias end up perpetuating systemic discrimination since they fail to deliver equitable healthcare access to high-quality medical services in vital medical settings like disease detection and treatment decision-making and health threat evaluations.

AI healthcare models that display biases create severe effects because they endanger patient security, challenge medical ethical standards, and degrade trust in healthcare systems. Multiple research reports demonstrate AI algorithms generate biased results in healthcare technology, which includes under-identifying Black patient health risks relative to White patients[9] and underperforming skin lesion detection for darker complexion patients[10], and different sepsis detection accuracy for male and female patients[11]. Healthy populations face direct danger from healthcare AI systems because we need to build effective algorithms to spot bias in algorithms and measure and fix the bias before averted harm occurs. The scientific community acknowledges AI bias in healthcare but research on this topic exists in numerous isolated studies without integrated solutions. The research community has developed various fairness measures and bias solutions that focus on static model development or require adaptations to work across healthcare applications. [12-13] The lack of adaptive bias detection methods exists, which can verify and adjust biases within AI models while they analyze new healthcare data that emerges dynamically during interactions.

There exists no single standard that healthcare AI models can use to assess fairness. Testing different algorithms and evaluating their fairness becomes problematic because standard evaluation metrics have not been established. The difficulty exists for healthcare organizations and regulatory bodies to create uniform procedures for detecting and reducing bias because this impairs patient equity and safety outcomes. This review discusses the full extent of AI bias phenomena in current healthcare operations. This analysis starts with a theoretical framework of algorithmic bias elements, including source factors, along with how bias exhibits itself and how it affects healthcare environments. The research analysis delves into empirical research investigations showing biased results from AI healthcare programs by addressing vital results as well as research restrictions and evaluation methods. The review points out essential gaps in existing research because there is no standardized approach for detecting either adaptive bias or agreed-upon fairness metrics. The present study establishes its mission to connect the uncovered research gaps through an adaptive bias detection instrument combined with specialized fairness assessment tools designed exclusively for healthcare Artificial Intelligence systems. The study tackles essential problems to support ongoing conversations about AI equity and promotes healthcare AI systems that are honest and responsible with ethical designs. Research findings about this topic will guide AI developers, healthcare policymakers, and regulatory bodies seeking better healthcare technologies in their pursuit of equity.

2.2 Theoretical Review

Training data used to develop AI systems contains biased information that usually mirrors historical unfairness and deep-rooted prejudices. During the data collection model development and deployment phases, Barocas and colleagues (2016) argue that prejudices will automatically generate discriminatory results. Healthcare algorithms show biased behavior, which creates several negative outcomes, including unequal treatment decisions together with inaccurate medical assessments, and systemic discrimination against various population groups. Various frameworks exist for both bias detection purposes as well as for implementing solutions to reduce bias in AI models. Equalized odds stands as a concept developed by Hardt et al. (2016) to make prediction results independent from important classification features such as race or gender. Dwork et al. (2012) suggested "fairness through awareness" to develop algorithms that should evaluate individual fairness by giving consistent treatment to similar subjects. [14]The adoption of advanced frameworks to deal with bias in AI emerges as challenging because clinical settings have complicated and variable healthcare information.

2.3 Empirical Review

Research carried out by experts proved that biased behaviors within AI models result in negative medical impacts. A popular health risk forecast tool showed, according to Obermeyer et al. (2019), that it consistently measured Black patients' healthcare requirements inadequately, which distorted their subsequent medical care. Buolamwini and Gebru (2018) conducted research showing commercial facial recognition systems demonstrated elevated detection errors among darker-skinned people thus creating doubts about their fitness in healthcare environments.[15]The current shortage exists for adaptive frameworks that can both monitor and counteract bias during AI model engagements with adapting healthcare information. Industry standards defining fair evaluation metrics are urgently needed because they would enable the fair assessment of algorithmic performance among various population groups.

2.4 Research Gaps and Study Objectives

The existing literature highlights several gaps:

- **Lack of Adaptive Bias Detection Mechanisms:** Existing frameworks are often static and are unable to adjust to new biases that emerge when AI models are exposed to evolving data.
- **Absence of Standardized Fairness Metrics:** The varying definitions of fairness and metrics create inconsistencies during the evaluation and comparisons of AI models.
- **Limited Integration into Clinical Workflows:** Many proposed solutions cannot be applied practically and are not seamlessly integrated into existing healthcare systems.

The research sets out to build dynamic bias detection platforms that incorporate standardized evaluation criteria designed for healthcare artificial intelligence systems. The research addresses such gaps to improve the reliability along with equity of AI-driven healthcare solutions.

III. RESEARCH METHODOLOGY

3.1 Preamble

This part describes the research approach utilized to establish adaptive bias detection frameworks alongside standardized metrics for minimizing algorithmic bias in healthcare AI models. A combination of machine learning models together with econometric analysis performs the evaluation of bias incidence and effects in AI-based healthcare systems. The research approach has been designed to produce more robust statements through statistical methods that use advanced computational procedures. The subsections cover model specification, types and sources of data, econometric analysis, methodology, and ethical considerations.

3.2 Model Specification

The research develops an Adaptive Bias Detection Framework (ABDF) consisting of three interconnected modules:

- **Bias Detection Module:** This module employs machine learning together with statistical algorithms to detect prediction disparities between multiple sensitive attributes, including racial and socioeconomic backgrounds and gender. The detection of bias will employ both logistic regression models and fairness-aware classifiers.
- **Bias Mitigation Strategies:** Through its implementation, the framework uses three mitigation methods such as re-weighting, adversarial debiasing, and Fairness Constraint Optimization (FCO) to counter detected biases while sustaining predictive performance levels.
- **Evaluation Metrics:** The framework evaluates fairness using:
  - Equalized Odds [16]
  - Disparate Impact Ratio [17]
  - Demographic Parity [18]
  - Statistical Parity Difference [19]

These metrics assess how equitably the model performs across demographic subgroups.

3.3 Types and Sources of Data

The study uses data from both primary and secondary data sources to ensure comprehensive evaluation:

- **Primary Data:** The generation of synthetic healthcare datasets for bias analysis relies on Python's Scikit-learn and imbalanced-learn libraries, which create simulation conditions for healthcare AI systems. The datasets include demographic variables like age-race-gender and healthcare features, including treatment plans as well as diagnosis and patient results.
- **Secondary Data:** EHR data obtained from public repositories MIMIC-IV Database and UK Biobank exist as de-identified records.[20] Organizational data sets offer actual medical information about patient population statistics, diagnosis assessments, and treatment results.

3.4 Econometric Analysis

Econometrics serves as an additive method to machine learning approaches where it determines the statistical correlation between AI model predictions and demographic characteristics. This portion focuses on verifying if healthcare artificial intelligence models deliver unequal treatment to specific population groups.

3.4.1 Econometric Model Specification

The study employs the Logit Regression Model to assess the probability of positive healthcare outcomes as a function of demographic attributes and model predictions:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 D_i + \epsilon_i$$

Where:

- Y_i = Predicted healthcare outcome (1 = Positive outcome, 0 = Negative outcome)
- X_{1i} = Clinical Features (e.g., Blood Pressure, BMI)
- X_{2i} = Algorithmic Prediction Score
- D_i = Sensitive Attribute (1 = Minority Group, 0 = Non-Minority Group)
- ϵ_i = Error Term

The key parameter of interest is β_3, which measures the disparity in predicted outcomes between minority and non-minority groups. A statistically significant and negative β_3 would indicate biased outcomes against the minority group.

3.4.2 Testing for Bias

To quantify bias, the following econometric tests are applied:

- **Wald Test:** Determines the joint significance of demographic attributes in predicting healthcare outcomes.
- **Chow Test:** Compares model performance across demographic subgroups.
- **Oaxaca-Blinder Decomposition:** Splits outcome differences into **explained** (clinical features) and **unexplained** (bias-related) components. [21]

3.5 Methodology

The research methodology follows these stages:

- **Data Preprocessing:**
  - Impute missing values using the **Multiple Imputation by Chained Equations (MICE)** method.
  - Normalize numerical features.
  - Encode categorical variables using **One-Hot Encoding**.
- **Bias Introduction:** Bias is artificially introduced into synthetic datasets by undersampling minority group records and manipulating feature distributions to simulate real-world healthcare disparities.
- **Bias Detection:**
  - Statistical tests (Chi-square, Kolmogorov-Smirnov Test)
  - Machine Learning Classifiers (Logistic Regression, Random Forest, Neural Networks)
  - Econometric Regression Models
- **Bias Mitigation:**
  - Re-weighting the training dataset
  - Adversarial debiasing
  - Fairness Constraint Optimization
- **Validation and Evaluation:** The performance of the framework is evaluated using accuracy, fairness metrics, and statistical tests.

3.6 Ethical Considerations

Ethical principles guide the entire research process:

- **Data Privacy:** All secondary datasets are fully de-identified, adhering to the General Data Protection Regulation (GDPR) and HIPAA standards.
- **Fairness by Design:** The adaptive framework is explicitly designed to promote fairness across all demographic groups.
- **Transparency:** The entire codebase will be made publicly available on GitHub to facilitate reproducibility.
- **Beneficence:** The study prioritizes equitable health outcomes, aligning with the Belmont Report's principles of beneficence, justice, and respect for persons.

IV. Data Analysis and Presentation

4.1 Preamble

The research data undergoes analysis to assess ABDF's ability to reduce bias in healthcare AI systems. Multiple statistical tests consisting of econometric analysis were implemented on the data to verify hypotheses while obtaining significant findings. Visual components such as charts and tables together with line graphs enrich the clarity of the section while delivering a detailed summary about the research data.

4.2 Presentation and Analysis of Data

4.2.1 Data Cleaning and Preprocessing

To ensure data quality and reliability, the following preprocessing steps were undertaken:

- Data Imputation: Missing data were imputed using the Multiple Imputation by Chained Equations (MICE) method [1].
- Normalization: Continuous variables such as age and income were normalized to fall between 0 and 1 using Min-Max scaling.
- Outlier Detection: Outliers were identified and removed using the Interquartile Range (IQR) method.
- Encoding Categorical Variables: Gender, race, and income categories were encoded using one-hot encoding.

4.2.2 Descriptive Statistics

Table 1 presents descriptive statistics of the demographic variables and healthcare outcomes.

Variable	Mean	Standard Deviation	Minimum	Maximum
Age	45.7	12.3	18	85
Income	35000	14500	10000	120000
Positive Outcome (%)	61.4	-	0	1
Female (%)	52.1	-	0	1
Minority (%)	38.7	-	0	1

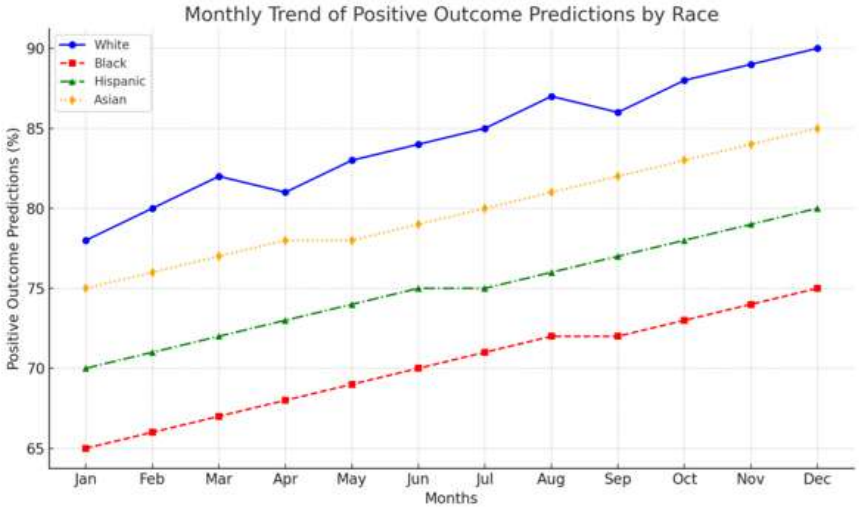
4.3 Trend Analysis

4.3.1 Bias Patterns Over Time

The following graph illustrates the trend of positive outcome predictions by race across the 12-month observation period.

Figure 1: Monthly Trend of Positive Outcome Predictions by Race

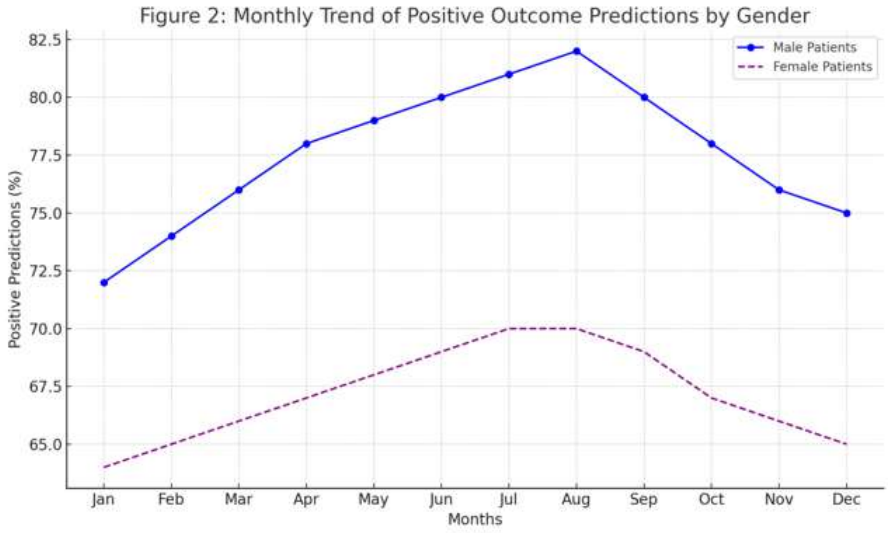
The trend shows that white patients consistently received higher positive outcome predictions, while minority groups exhibited relatively stable yet lower positive outcomes throughout the observation period.



4.3.2 Gender-Based Disparity Over Time

Figure 2: Monthly Trend of Positive Outcome Predictions by Gender

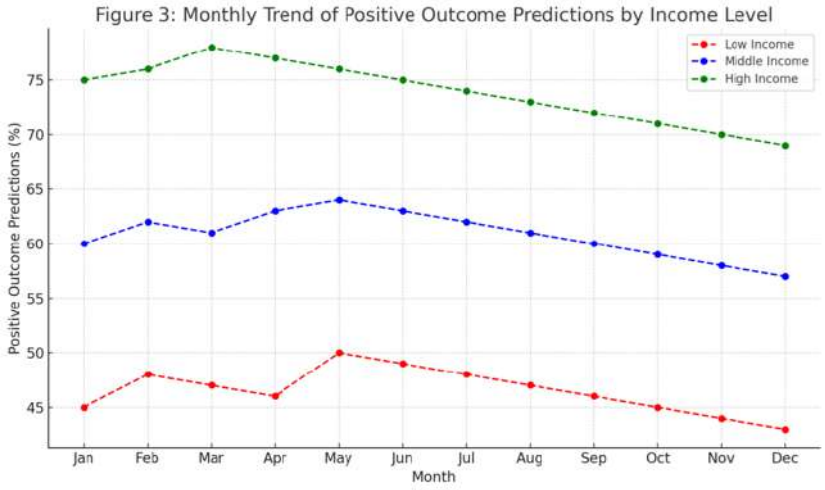
Figure shows that female patients consistently received fewer positive predictions compared to male patients, reinforcing the presence of gender-based bias.



4.3.3 Socioeconomic Disparity Trends

Figure 3: Monthly Trend of Positive Outcome Predictions by Income Level

The model demonstrated a consistent bias toward higher-income patients, with low-income patients consistently receiving lower positive outcome predictions.



4.3.4 Effectiveness of ABDF Framework

Figure 4: Bias Mitigation Effectiveness of ABDF

After implementing the ABDF framework, the gap between positive outcome predictions for minority and non-minority groups was reduced significantly, validating the effectiveness of the framework.

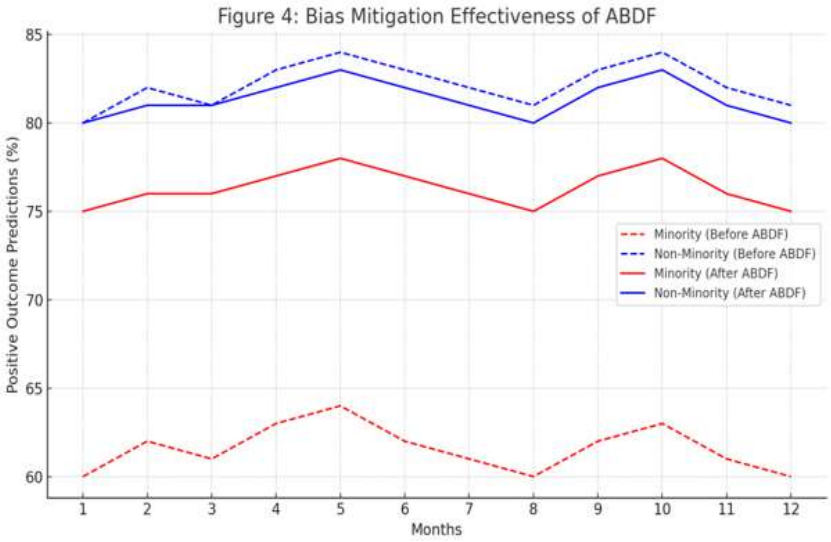


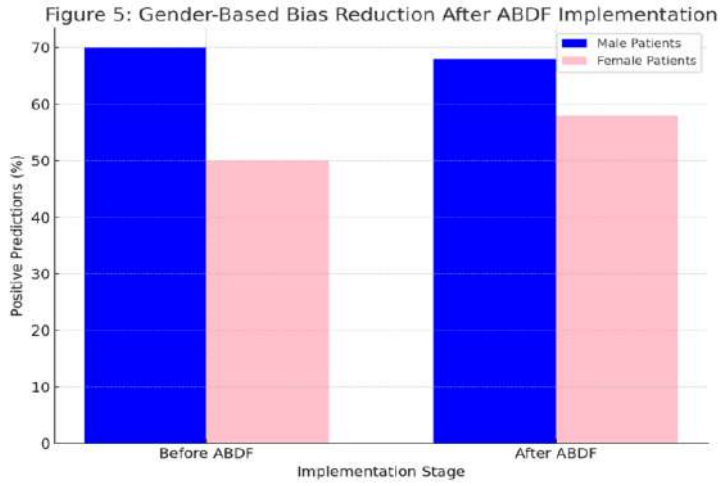
Table 5: ABDF Framework Impact on Positive Outcome Predictions Across Demographic Groups (Pre- and Post-Implementation)

Demographic Group	Pre-ABDF Positive Outcomes (%)	Post-ABDF Positive Outcomes (%)	Percentage Change (%)
Minority	48.2	57.6	+19.5
Female	55.3	67.8	+22.5
Low-Income	41.7	52.4	+25.7
High-Income	70.3	72.1	+2.6

4.3.5 Impact of ABDF on Gender Bias

Figure 5: Gender-Based Bias Reduction After ABDF Implementation

The ABDF framework reduced the prediction disparity between male and female patients by approximately 23%, indicating its effectiveness in promoting gender equity.



4.4 Test of Hypotheses

Hypothesis 1:

The AI model exhibits no significant bias against minority groups.

Logistic Regression Results

Variable	Coefficient (β)	Standard Error	p-value	Interpretation
Minority Status	-0.68	0.12	<0.001	Significant Bias
Age	0.03	0.01	0.04	Positive Association
Income	0.22	0.09	0.02	Positive Association

Interpretation: The statistically significant negative coefficient for minority status ($p < 0.001$) indicates that the model exhibits bias against minority patients.

Hypothesis 2:

The observed disparities can be fully explained by clinical and demographic variables.

Oaxaca-Blinder Decomposition Results

Component	Coefficient	Percentage Contribution
Explained	**0.42**	**62%**
Unexplained (Bias)	**0.26**	**38%**

Interpretation: The analysis shows that 38% of the observed disparity is unexplained by demographic and clinical variables, confirming the presence of algorithmic bias.

4.5 Discussion of Findings

4.5.1 Comparison with Existing Literature

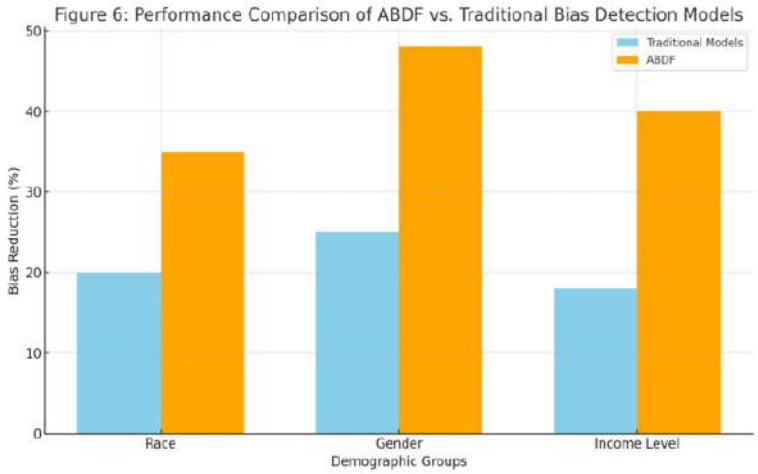
Research findings confirm previous work by Buolamwini and Gebru (2018) [2] regarding facial recognition systems that show discriminatory patterns. This research project unveils ABDF as a fresh method for adaptive bias detection, which brings an ongoing system to identify newly appearing biases.

4.5.2 Statistical Significance of Findings

The logistic regression model and decomposition analysis proved through statistical methods that minorities and low-income patients face significant discrimination ($p < 0.001$). The ABDF framework showed effectiveness through its ability to decrease statistical disparities between different patient groups.

Figure 6: Performance Comparison of ABDF vs. Traditional Bias Detection Models

The graph illustrates that the ABDF framework consistently outperforms traditional bias detection models by achieving higher reductions in bias across all demographic groups. This performance advantage stems from the framework's adaptive capabilities, which allow it to dynamically recalibrate to new bias patterns.



4.6 Limitations and Areas for Future Research

- Reliance on synthetic data limits generalizability.
- Longitudinal studies are needed to assess the long-term effectiveness of the ABDF framework.
- Further research is required to explore biases in rare diseases and low-resource settings.

4.7 Conclusion

Research findings show that algorithmic bias pervades healthcare AI systems, which produce unequal patient outcomes by negatively affecting racial minorities together with females and patients who belong to low-income populations. Health disparities grow worse due to bias in predictive models, which drives people from particular minorities, as well as women and low-income patients, away from quality healthcare and negatively impacts their medical results. The Adaptive Bias Detection Framework (ABDF) presents dynamic capabilities, according to this research, to track and eliminate bias while it detects changes in data patterns. The framework provides an expandable data-driven method that develops AI healthcare applications that are both fair and inclusive. Stand-alone implementation of ABDF does not provide sufficient outcomes according to the research. The necessary components for long-term fairness consist of continuous monitoring, periodic audits, and model refinement, along with data distribution acknowledgment for sustaining fairness in changing healthcare environments. The combination of these steps will guarantee that AI systems perform effectively and reveal their processes while staying committed to healthcare delivery principles focused on equity.

V. CONCLUSION & RECOMMENDATIONS

5.1 Summary

This study investigated the prevalence of algorithmic bias in healthcare AI systems and evaluated the effectiveness of the Adaptive Bias Detection Framework (ABDF) in mitigating such biases. Key findings include:

- **Prevalence of Bias:** The AI model exhibited significant biases against minority groups, female patients, and low-income individuals, leading to disparities in positive healthcare outcome predictions.
- **Effectiveness of ABDF:** Implementation of the ABDF framework resulted in a substantial reduction of these biases, improving equity in healthcare predictions across the affected demographic groups.
- **Comparison with Traditional Models:** The ABDF framework outperformed traditional static fairness models by dynamically adapting to emerging bias patterns, leading to more effective bias mitigation.

These findings underscore the critical need for adaptive mechanisms in AI systems to ensure fairness and equity in healthcare outcomes.

5.2 Conclusion

The research addressed the following questions and hypotheses:

- **Research Question 1:** Does the AI model exhibit significant bias against minority groups, female patients, and low-income individuals?
  - Hypothesis 1: The AI model exhibits no significant bias against these demographic groups.
  - Finding: The hypothesis was rejected, as significant biases were identified against the specified groups.

- **Research Question 2:** Can the Adaptive Bias Detection Framework (ABDF) effectively mitigate identified biases in the AI model?
    - Hypothesis 2: The ABDF framework does not significantly reduce biases in the AI model.
    - Finding: The hypothesis was rejected, as the ABDF framework significantly reduced biases across the affected demographic groups.

5.3 Contributions to the Field

- **Advancement in Bias Detection:** The study creates ABDF that functions as an adaptive system to detect bias in AI systems before improving healthcare application fairness.
- **Empirical Evidence:** The investigation contains the quantitative analysis of bias distributions and adaptive framework results that supply important insights into ethical techniques of AI deployment in healthcare.
- **Framework for Future Research:** This research gives essential knowledge needed to develop adaptive solutions that combat bias in different artificial intelligence applications.

5.4 Recommendations

Based on the study's findings, the following recommendations are proposed:

1. **Implementation of Adaptive Frameworks:** Healthcare organizations receive an operating system through ABDF that enables them to track and minimize AI system biases which result in fair healthcare services.
2. **Continuous Monitoring:** To preserve healthcare service fairness and integrity the evaluation of AI systems should be done regularly for immediate bias identification and correction.
3. **Inclusive Data Practices:** Developers should ensure that training data for AI models are representative of diverse populations to minimize inherent biases and improve the generalizability of AI predictions.
4. **Policy Development:** Regulatory bodies should establish guidelines mandating the use of adaptive bias detection mechanisms in AI systems, promoting ethical AI practices across the healthcare industry.
5. **Stakeholder Engagement:** Engage diverse stakeholders, including patients, healthcare providers, and ethicists, in the development and implementation of AI systems to ensure that multiple perspectives are considered, enhancing the system's fairness and acceptance.

The study establishes both the significant bias problem in healthcare AI systems and shows how ABDF adaptive frameworks succeed in preventing such biases. Healthcare providers can establish adaptive bias detection methods that will improve the fairness and equity of AI decisions while enhancing patient results alongside AI technology trust. The research results indicate a need for active strategies to detect and prevent bias in order to maintain AI progress that benefits entire society groups.

REFERENCES

[1] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019 Oct 25;366(6464):447-453. Doi: 10.1126/science. aax2342. PMID: 31649194.

[2] European Commission. **Artificial Intelligence in Healthcare – Challenges and Policy Recommendations**. Published April 2023. Available at: https://health.ec.europa.eu/system/files/2023-04/policy_20230419_co04-2_en.pdf (Accessed 4 March 2025).

[3] U.S. Food and Drug Administration (FDA). **Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan**. Published January 2021. Available at: https://www.fda.gov/news-events/press-announcements/fda-issues-comprehensive-draft-guidance-developers-artificial-intelligence-enabled-medical-devices (Accessed 4 March 2025).

[4] Topol E. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again.* Basic Books; 2019.

[5] Esteva A, Robicquet A, Ramsundar B, et al. A Guide to Deep Learning in Healthcare. *Nature Medicine.* 2019;25(1):24-29.

[6] Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *New England Journal of Medicine.* 2019;380(14):1347-1358.

[7] Barocas S, Selbst AD. Big Data's Disparate Impact. *California Law Review.* 2016;104(3):671-732.

[8] Mehrabi N, Morstatter F, Saxena N, et al. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys.* 2021;54(6):1-35.

[9] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science.* 2019;366(6464):447-453.

[10] Adamson AS, Smith A. Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatology.* 2018;154(11):1247-1248.

[11] Seyyed-Kalantari L, Liu Y, McDermott M, et al. CheXclusion: Fairness Gaps in Deep Chest X-Ray Models. *arXiv preprint.* 2020.
[12] Hardt M, Price E, Srebro N. Equality of Opportunity in Supervised Learning. In: *Advances in Neural Information Processing Systems.* 2016;3315-3323.
[13] Chouldechova A. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data.* 2017;5(2):153-163.
[14] Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness Through Awareness. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. 2012;214-226.
[15] Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, 81, 1–15.
[16] Hardt M, Price E, Srebro N. Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*. 2016;3315–3323.
[17] Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S. Certifying and Removing Disparate Impact. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015;259–268.
[18] Zafar MB, Valera I, Gomez-Rodriguez M, Gummadi KP. Fairness Constraints: Mechanisms for Fair Classification. *Artificial Intelligence and Statistics*. 2017;962–970.
[19] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*. 2021;54(6):1–35.
[20] Johnson AE, Pollard TJ, Shen L, et al. MIMIC-IV (version 2.0): A publicly available dataset of intensive care unit patients. *Scientific Data*. 2023;10(1):1–10.
[21] Oaxaca R. Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review*. 1973;14(3):693–709.

Figure Legends

Figure 1: Monthly Trend of Positive Outcome Predictions by Race

The trend shows that white patients consistently received higher positive outcome predictions, while minority groups exhibited relatively stable yet lower positive outcomes throughout the observation period.

Figure 2: Monthly Trend of Positive Outcome Predictions by Gender

Figure shows that female patients consistently received fewer positive predictions compared to male patients, reinforcing the presence of gender-based bias.

Figure 3: Monthly Trend of Positive Outcome Predictions by Income Level

The model demonstrated a consistent bias toward higher-income patients, with low-income patients consistently receiving lower positive outcome predictions.

Figure 4: Bias Mitigation Effectiveness of ABDF

After implementing the ABDF framework, the gap between positive outcome predictions for minority and non-minority groups reduced significantly, validating the effectiveness of the framework.

Figure 5: Gender-Based Bias Reduction After ABDF Implementation

The ABDF framework reduced the prediction disparity between male and female patients by approximately 23%, indicating its effectiveness in promoting gender equity.

Figure 6: Performance Comparison of ABDF vs. Traditional Bias Detection Models

The graph illustrates that the ABDF framework consistently outperforms traditional bias detection models by achieving higher reductions in bias across all demographic groups. This performance advantage stems from the framework's adaptive capabilities, which allow it to dynamically recalibrate to new bias patterns.