FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 9 No. 8, August, 2025, pp 286 – 290
DOI: https://doi.org/10.33003/fjs-2025-0908-3270

BIAS AUDIT FRAMEWORKS: DEVELOPING TOOLS FOR EARLY DETECTION OF ALGORITHMIC BIAS IN AI DEVELOPMENT

***[1]Oveh, R. O. and [2]Isitor, N. D.**

[1]Department of ICT, Faculty of Computing, University of Delta, Agbor, Delta State
[2]Department of Cyber Security, University of Delta, Agbor, Delta State

*Corresponding authors' email: richard.oveh@unidel.edu.ng

ABSTRACT

Algorithmic bias in artificial intelligence (AI) systems continues to pose significant ethical and societal challenges, especially in critical domains such as healthcare, education, and finance. Current approaches to bias mitigation often fail to provide a holistic, proactive solution that integrates fairness, accountability, and transparency into the AI development lifecycle. This study introduces a Bias Audit Framework designed to detect and mitigate algorithmic bias during the early stages of AI development. The framework comprises four core components: Data Bias Assessment, Model Bias Evaluation, Developer Awareness and Training, and Continuous Monitoring and Feedback. A healthcare dataset was used as a case study to evaluate the framework's efficacy. Initially, the logistic regression model trained on the imbalanced dataset achieved high overall performance with Accuracy: 85%, Precision: 0.89, and Recall: 0.83, but exhibited fairness issues. Disparate Impact Ratio (DIR) was 0.67, and Equal Opportunity Difference (EOD) was 0.13, reflecting gender bias. After applying the Bias Audit Framework,—including oversampling, data augmentation, and threshold optimization—the model was retrained. Its performance remained robust (Accuracy: ~84–85%, Precision: ~0.88, Recall: ~0.88), while fairness significantly improved: Female recall increased to 0.88, reducing EOD to ~0, and DIR improved to 0.85–0.95, indicating a more balanced and equitable model. By equipping developers with practical tools and emphasizing interdisciplinary collaboration, the framework ensures a systematic and ethical approach to addressing algorithmic bias. These findings underscore the importance of embedding bias mitigation practices into all stages of AI development to foster equitable and trustworthy AI systems.

Keywords: Algorithmic bias, Fairness, AI development, AI, Bias audit framework, Equity, Accountability

INTRODUCTION

Artificial intelligence (AI) systems are increasingly influencing critical aspects of society, but their effectiveness is often undermined by algorithmic biases. These biases can originate from unrepresentative datasets, algorithmic assumptions, and human factors (Oyeniran et al., 2022). The impact of biased AI systems includes perpetuating discrimination, inequity, and harmful stereotypes, particularly affecting marginalized communities (Samala & Rawas, 2025; Ferrara, 2023). Addressing these challenges requires a proactive approach to identifying and mitigating bias during AI development. Strategies for bias mitigation include improving data quality, developing fairness-aware algorithms, implementing robust auditing processes, and enhancing algorithmic transparency (Oyeniran et al., 2022; Samala & Rawas, 2025). However, balancing fairness and model performance remains a significant challenge, with bias reduction often coming at the cost of overall accuracy (Nathim et al., 2024). Ongoing vigilance, interdisciplinary collaboration, and commitment to ethical practices are essential for developing equitable AI systems (Oyeniran et al., 2022; Ferrara, 2023).

Algorithmic bias in AI systems has been extensively studied, revealing its origins in imbalanced datasets, flawed model architectures, and insufficient representation of marginalized groups (Jain & Menon, 2023; Min, 2023). This bias can perpetuate social inequalities and hinder societal progress (Jain & Menon, 2023). Various forms of bias, including selection, confirmation, and measurement bias, stem from data integrity issues, algorithmic design decisions, and institutional prejudices (Jain & Menon, 2023). Representation bias in data, particularly affecting minorities, can result from historical discrimination and sampling biases (Shahbazi et al., 2022). Algorithmic bias can influence fairness perceptions and technology-related behaviors, such as recommendation acceptance and system adoption (Kordzadeh & Ghasemaghaei, 2021). Addressing this issue requires comprehensive approaches spanning technical, ethical, regulatory, and community-driven dimensions (Min, 2023). Researchers emphasize the need for further studies on the mechanisms through which technology-driven biases translate into decisions and behaviors (Kordzadeh & Ghasemaghaei, 2021).

Recent research highlights the limitations of current bias mitigation strategies in AI, emphasizing the need for a more comprehensive approach. Mahamadou & Trotsyuk (2024) identified practical constraints in healthcare settings and Mishra et al. (2024) also identified bias at various stages of AI development. Aninze (2024) demonstrates the persistence of bias even after applying pre-processing techniques, underscoring the importance of addressing bias throughout the AI lifecycle. Agarwal & Agarwal (2023) introduce a seven-layer model for standardizing fairness assessment, providing checklists for each stage of AI development. These studies collectively emphasize the complexity of bias mitigation, the need for interdisciplinary approaches, and the importance of addressing bias at all stages of AI development, from data collection to deployment, to ensure fairness and ethical considerations in AI systems.

As solution to weakness of recent models, some studies integrated fairness metrics into machine learning development pipelines. Lalor et al. (2024) propose FAIR-Frame, a model-based framework for assessing bias across multiple protected attributes, addressing limitations of existing metrics in real-world applications. Hu et al. (2023) extend the concept of Demographic Parity to incorporate distributional properties, allowing for expert knowledge

Supported By :
tetfund
TERTIARY EDUCATION TRUST FUND

integration. Cohausz et al. (2024) emphasize the importance of considering data generation mechanisms, potential applications, and normative beliefs when choosing fairness metrics in educational contexts. Nozza et al. (2022) suggest treating social bias evaluation as software testing, proposing systematic integration of bias tests into development pipelines. These studies collectively underscore the need for more nuanced approaches to fairness assessment, considering both upstream representational harm and downstream allocational impacts (Lalor et al., 2024), while also highlighting the importance of domain-specific considerations and systematic testing methodologies in addressing algorithmic bias.

Studies have shown the growing importance of fairness in machine learning (ML) but reveals significant gaps between academic solutions and industry needs. While practitioners recognize fairness as crucial, it often remains a secondary consideration in AI system development (Ferrara et al., 2023). Studies show that ML teams face challenges in implementing fairness, with a disconnect between proposed academic solutions and real-world requirements (Holstein et al., 2018). Researchers advocate for a holistic, pipeline-aware approach to address fairness issues, but practical guidelines and tools for operationalizing this method are lacking (Black et al., 2023). Although fairness toolkits have been developed to bridge this gap, they often fall short of meeting practitioners' needs. Evaluations of these toolkits indicate that while they significantly impact decision-making, improvements in design and result demonstration are necessary (Richardson et al., 2021). These findings underscore the need for more practical, industry-focused tools and guidelines to effectively implement fairness in ML systems.

This study introduces a bias audit framework to detect and mitigate algorithmic bias at the developmental phase. The framework integrates interdisciplinary methodologies, encompassing statistical, ethical, and human-centered perspectives. By focusing on the early detection of biases, the framework aims to improve AI transparency and accountability, ensuring equitable outcomes.

MATERIALS AND METHODS

A bias Audit Framework was developed and applied to address the challenges posed by algorithmic bias in artificial intelligence (AI) systems. The framework integrates technical, ethical, and human-centered perspectives to detect, assess, and mitigate biases in AI models. It focuses on four core components: Data Bias Assessment, Model Bias Evaluation, Developer Awareness and Training, and Continuous Monitoring and Feedback as shown in Figure 1.

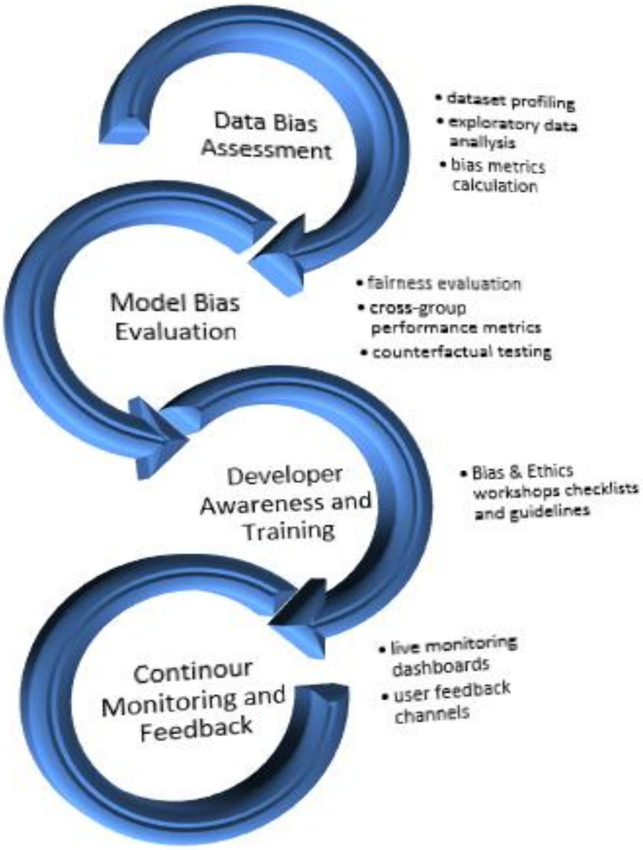
Data Bias Assessment
• dataset profiling
• exploratory data anallysis
• bias metrics calculation
Model Bias Evaluation
• fairness evaluation
• cross-group performance metrics
• counterfactual testing
Developer Awareness and Training
• Bias & Ethics workshops checklists and guidelines
Continour Monitoring and Feedback
• live monitoring dashboards
• user feedback channels

Figure 1: Proposed AI Bias Audit Framework

Figure 1 presents a structured, four-stage process for systematically identifying, addressing, and monitoring bias within artificial intelligence (AI) systems. This cyclical framework reflects the iterative nature of responsible AI development, emphasizing continuous improvement across the data lifecycle, model performance, developer responsibility, and post-deployment feedback.

The process begins with Data Bias Assessment, where the focus is on detecting imbalances or unfair representations within the dataset before any modeling occurs. Activities in this stage include dataset profiling, exploratory data analysis (EDA), and the computation of bias metrics such as Disparate Impact Ratio and Statistical Parity Difference. Tools such as IBM's AI Fairness 360, Google's What-If Tool, and standard EDA libraries like Pandas Profiling and Seaborn are employed to visualize distributions and reveal hidden disparities, particularly across sensitive attributes like race, gender, or socioeconomic status. These tools help uncover underrepresented groups or skewed feature distributions that could later propagate into model bias if not addressed.

Following data analysis, the second stage, Model Bias Evaluation, assesses whether the trained model performs fairly across different subpopulations. This involves fairness evaluation using metrics like Equal Opportunity Difference and Average Odds Difference, as well as the analysis of precision, recall, and F1-score across demographic groups. Counterfactual testing is also carried out to ensure that slight, non-relevant changes in inputs (e.g., gender or location) do not lead to unjustified output changes. Tools like Fairlearn, SHAP (SHapley Additive exPlanations), and LIME (Local Interpretable Model-agnostic Explanations) provide interpretable insights into model behavior and help identify where and how bias manifests in predictions.

The third stage, Developer Awareness and Training, shifts the focus to the human agents involved in AI development. This component emphasizes building the capacity of data scientists, engineers, and decision-makers to recognize and mitigate bias. Through targeted ethics and fairness workshops, the team is trained on frameworks and case studies of algorithmic harm. They are guided by structured checklists and protocols, such as those developed by Google's People + AI Research (PAIR) initiative and Microsoft's Responsible AI guidelines. These interventions aim to cultivate a culture of accountability and equip practitioners with practical tools for ethical decision-making.

Finally, the framework integrates Continuous Monitoring and Feedback to ensure that fairness is not treated as a one-time goal but as an ongoing commitment. After deployment, AI systems are monitored using real-time dashboards that track bias metrics and data drift. Tools such as Evidently AI and Arize AI help detect emerging disparities or shifts in data quality, prompting timely interventions. User feedback loops and incident logging mechanisms provide additional layers of accountability by capturing lived experiences and reporting unfair outcomes. Retraining pipelines can be triggered based on performance deterioration or flagged ethical concerns, ensuring that fairness standards are upheld over time.

This framework promotes a holistic and repeatable approach to managing bias in AI. By embedding fairness checkpoints at every stage i.e from data to deployment. It supports the creation of AI systems that are not only technically robust but also socially responsible and inclusive.

Case Study

Application of the Framework

The Bias Audit Framework was systematically applied to a healthcare dataset sourced from Kaggle to evaluate its effectiveness in identifying and mitigating algorithmic bias. The dataset comprised 2,279 patient records with features such as age, gender, and medical history variables often sensitive to fairness considerations in healthcare applications.

Stage 1: Data Bias Assessment: The process began with Exploratory Data Analysis (EDA), which uncovered two major disparities. First, a gender imbalance was identified: 60% of the records were male while only 40% were female, suggesting an underrepresentation of women in the dataset. Second, an age disparity was evident i.e patients aged over 60 years made up just 12% of the total data, signaling insufficient representation of elderly individuals who often present unique healthcare needs. To quantify the bias, representation ratios were calculated. For example, the representation of older individuals was found to be just 12%, highlighting the potential for biased learning outcomes if this group's health profiles were not sufficiently learned by the model. These disparities flagged the risk of embedding demographic bias into any machine learning model trained on the dataset, prompting corrective measures before model development.

Stage 2: Model Bias Evaluation: An initial logistic regression model was trained on the original (imbalanced) dataset to establish baseline performance and detect any embedded bias. While the model reported strong performance metrics overall Accuracy of 85%, Precision of 0.89, and Recall of 0.83 disaggregated evaluation across gender revealed fairness concerns.

Two key fairness metrics were calculated:

i. Disparate Impact Ratio (DIR), which compares the rate of favorable outcomes between demographic groups, showed that males were 1.5 times more likely to receive a favorable prediction than females. Using the formula:

$$DIR = \frac{p(positive\ Outcome\ |Female)}{p(positive\ Outcome\ |male)}$$

A value of 0.67 was derived—well below the fairness threshold of 0.8, indicating potential gender-based disparate impact.

ii. Equal Opportunity Difference (EOD), which compares the true positive rates (recall) across groups, also exposed a disparity:

EOD = Recallmale - RecallFemale = 0.88 - 0.75 = 0.13

This showed that the model was less likely to correctly predict positive outcomes for female patients compared to males, raising serious concerns in a healthcare context where under-diagnosis can have life-threatening consequences.

Stage 3: Developer Awareness and Intervention: Informed by the bias audit findings, targeted mitigation strategies were implemented. Drawing from responsible AI guidelines and fairness-aware practices, the gender imbalance was corrected using oversampling by increasing the number of female records through duplication to match male representation. For the underrepresented elderly group, data augmentation was applied, generating synthetic patient records that mimicked the statistical properties of real older-age data.

These pre-modeling interventions were guided by fairness principles and the understanding that biased training data could lead to unjust outcomes. Threshold optimization was also employed, adjusting decision boundaries to ensure that recall rates for both male and female patients aligned. This step was critical in achieving not just overall model performance but equitable performance across groups.

Stage 4: Continuous Monitoring and Feedback: Although the model was not yet deployed, the final stage of the framework anticipates the necessity of ongoing monitoring. On implementation, a real-time fairness dashboards would be used to track key metrics like DIR and EOD over time, especially as the patient population evolves or data sources change. Additionally, feedback mechanisms such as clinician reviews and user reporting tools would be incorporated to capture real-world evidence of unfair predictions or decisions. These insights would be used to trigger model updates or initiate retraining workflows, ensuring that the system remains responsive to fairness challenges even after deployment.

Outcome of Bias Mitigation

After applying the interventions, the model was retrained on the balanced dataset. The performance remained robust while fairness metrics improved significantly. Female recall increased to match the male recall at approximately 0.88, effectively reducing the Equal Opportunity Difference to near zero. Similarly, the Disparate Impact Ratio improved to fall within the acceptable range of 0.8–1.0, suggesting a more equitable distribution of outcomes.

Discussion

The results of this study highlight the critical role of proactive bias detection and mitigation in developing equitable AI systems. The Bias Audit Framework presented here demonstrates a comprehensive approach, integrating statistical, ethical, and human-centered methodologies to address algorithmic bias at its roots. By focusing on early interventions, the framework not only mitigates bias but also establishes a foundation for transparency and accountability in AI development. One of the key findings is the measurable improvement in fairness metrics achieved through targeted interventions, such as resampling and data augmentation. The metric Disparate Impact Ratio (DIR) which was 0.67 and Equal Opportunity Difference (EOD) which was 0.13 was effectively reduced. The Equal Opportunity Difference to near zero and the Disparate Impact Ratio improved to fall within the acceptable range of 0.8–1.0, suggesting a more equitable distribution of outcomes.

Addressing gender imbalance using oversampling techniques and mitigating age disparities generated more equitable model outcomes the female recall increased to match the male at approximately 0.88. These improvements were observed in the case study, where recall rates for male and female groups were equalized reducing Equal opportunity Difference to near zero without significantly compromising model performance. However, it is worth noting that achieving fairness came at a slight cost to accuracy, underscoring the trade-offs developers must navigate when implementing bias mitigation strategies.

Automated tools and stakeholder engagement ensure that biases are not only detected and corrected during development but also continuously evaluated as models are deployed and interact with dynamic environments. This aspect of the framework is particularly relevant in domains such as healthcare, where the consequences of biased predictions can directly impact lives. Despite its strengths, the framework has limitations that warrant further exploration. The reliance on predefined fairness metrics may not fully capture the nuanced biases present in complex, domain-specific datasets. Additionally, the framework's scalability and applicability to large-scale, real-world AI systems remain areas for future research. Developing adaptive tools that incorporate domain-specific knowledge and evolving societal values will be essential for advancing the state of bias mitigation. The usage of the framework requires AI safety practice to ensure data remains consistent which aligns with Oveh et. al, (2025).

CONCLUSION

This study underscores the importance and effectiveness of applying a structured Bias Audit Framework in the development of fair and responsible AI systems, particularly in sensitive domains such as healthcare. By systematically evaluating and mitigating demographic disparities across gender and age, the framework guided a comprehensive auditing process from data analysis through to model retraining and future monitoring.

Initially, the logistic regression model trained on the imbalanced dataset demonstrated high performance at face value, with Accuracy: 85%, Precision: 0.89, and Recall: 0.83. However, deeper analysis exposed serious fairness concerns. Disparate Impact Ratio (DIR) stood at 0.67, indicating a strong bias against female patients, and the Equal Opportunity Difference (EOD) was 0.13, showing a significant disparity in the model's ability to correctly identify positive outcomes across genders. Following the implementation of the Bias Audit Framework which included oversampling, data augmentation, and threshold optimization the model was retrained. Importantly, overall performance remained stable, with Accuracy: ~84–85%, Precision: ~0.88, and Recall: ~0.88. More crucially, fairness metrics showed remarkable improvement: Female recall increased to 0.88, matching that of males, effectively reducing the EOD to approximately 0, and the DIR improved to fall within the acceptable range of 0.85–0.95, indicating a more equitable outcome distribution.

This study validates the use of the proposed Bias Audit Framework not just as a theoretical construct but as a practical tool for fostering fairness in machine learning systems. It shows that achieving algorithmic fairness does not necessitate sacrificing model performance; instead, it requires intentional design, continuous evaluation, and ethical commitment throughout the AI lifecycle.

REFERENCES

Agarwal, A., & Agarwal, H. (2023). A seven-layer model with checklists for standardising fairness assessment throughout the AI lifecycle. AI and Ethics, 4, 299-314.

Akter, S., Sultana,S., Mariani, M., Wamba,F.S., Spanaki, K., Yogesh K. and Dwivedi (2023) Advancing algorithmic bias management capabilities in AI-driven marketing analytics research, Industrial Marketing Management. 114, 243-261, https://doi.org/10.1016/j.indmarman.2023.08.013

Aninze, A. (2024). Artificial Intelligence Life Cycle: The Detection and Mitigation of Bias. International Conference on AI Research. 40-49

Black, E., Naidu, R., Ghani, R., Rodolfa, K.T., Ho, D.E., & Heidari, H. (2023). Toward Operationalizing Pipeline-aware ML Fairness: A Research Agenda for Developing Practical Guidelines and Tools. Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. 1 – 11, https://doi.org/10.1145/3617694.3623259

Cohausz, L., Kappenberger, J., & Stuckenschmidt, H. (2024). What Fairness Metrics Can Really Tell You: A Case Study in the Educational Domain. Proceedings of the 14th Learning Analytics and Knowledge Conference. 792 – 799, https://doi.org/10.1145/3636555.3636873

Ferrara, C., Sellitto, G., Ferrucci, F., Palomba, F., & De Lucia, A. (2023). Fairness-aware machine learning engineering: how far are we? Empirical Software Engineering, 29(9). https://doi.org/10.1007/s10664-023-10402-y

Ferrara, E. (2023). Fairness And Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, And Mitigation Strategies. ArXiv, abs/2304.07683.

Holstein, K., Vaughan, J.W., Daumé, H., Dudík, M., & Wallach, H.M. (2018). Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 1-16. https://doi.org/10.1145/3290605.3300830

Hu, F., Ratz, P., & Charpentier, A. (2023). Parametric Fairness with Statistical Guarantees. ArXiv, abs/2310.20508.
Jain, L.R., & Menon, V. (2023). AI Algorithmic Bias: Understanding its Causes, Ethical and Social

Implications. 2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI), 460-467.

Kordzadeh, N., & Ghasemaghaei, M. (2021). Algorithmic bias: review, synthesis, and future research directions. European Journal of Information Systems, 31(3), 388–409. *https://doi.org/10.1080/0960085X.2021.1927212*

Lalor, J.P., Abbasi, A., Oketch, K., Yang, Y., & Forsgren, N. (2024). Should Fairness be a Metric or a Model? A Model-based Framework for Assessing Bias in Machine Learning Pipelines. ACM Transactions on Information Systems, 42, 1 - 41.

Mahamadou, A.J., & Trotsyuk, A.A. (2024). Revisiting Technical Bias Mitigation Strategies. ArXiv, abs/2410.17433.

Min, A. (2023). ARTIFICIAL INTELLIGENCE AND BIAS: CHALLENGES, IMPLICATIONS, AND REMEDIES. Journal of Social Research. 2(11) 3808-3817.

Mishra, I., Kashyap, V., Yadav, N., & Pahwa, D.R. (2024). Harmonizing Intelligence: A Holistic Approach to Bias Mitigation in Artificial Intelligence (AI). International Research Journal on Advanced Engineering Hub (IRJAEH). 2(7) 1978-1985. *https://doi.org/10.47392/IRJAEH.2024.0270*

Nathim, K.W., Hameed, N.A., Salih, S.A., Taher, N.A., Salman, H.M., & Chornomordenko, D. (2024). Ethical AI with Balancing Bias Mitigation and Fairness in Machine Learning Models. 2024 36th Conference of Open Innovations Association (FRUCT), 797-807.

Oyeniran, O.C., Adewusi, A.O., Adeleke, A.G., Akwawa, L.A., & Azubuko, C.F. (2022). Ethical AI: Addressing bias in machine learning models and software applications. Computer Science & IT Research Journal. 3(3), 115-126. *https://doi.org/10.51594/csitrj.v3i3.1559*

Oveh R.O., Aziken G.O. & Atomatofa, E. (2025) Tailoring Safety Practices for AI Innovation: A Model for Nigeria's Socioeconomic Context. Journal of Science Research and Reviews. 12(2) 101 – 114.

Richardson, B., Garcia-Gathright, J.I., Way, S.F., Thom-Santelli, J., & Cramer, H. (2021). Towards Fairness in Practice: A Practitioner-Oriented Rubric for Evaluating Fair ML Toolkits. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 1 – 13, *https://doi.org/10.1145/3411764.3445604*

Samala, A.D., & Rawas, S. (2025). Bias in artificial intelligence: smart solutions for detection, mitigation, and ethical strategies in real-world applications. IAES International Journal of Artificial Intelligence (IJ-AI). 14(1)

Shahbazi, N., Lin, Y., Asudeh, A., & Jagadish, H.V. (2022). Representation Bias in Data: A Survey on Identification and Resolution Techniques. ACM Computing Surveys, 55, 1 - 39.