International Journal of Information Technology Research and Development (IJITRD)
Volume 5 Issue 2, July – December (2024), pp. 5-11.

MITIGATING ALGORITHMIC BIAS IN CLINICAL AI SYSTEMS THROUGH FAIRNESS-AWARE TRAINING OBJECTIVES AND EQUITABLE MODEL CALIBRATION TECHNIQUES

Tharunika Rajendiran,
USA.

Abstract

The integration of artificial intelligence (AI) into clinical decision-making has the potential to improve diagnostic accuracy, treatment recommendations, and patient outcomes. However, the increasing reliance on machine learning (ML) in healthcare has revealed deep-rooted algorithmic biases that may exacerbate health disparities across demographic groups. This paper explores strategies to mitigate such biases through fairness-aware training objectives and post-hoc calibration techniques. We propose a framework combining group fairness constraints during training and recalibration methods to enhance equity across ethnic, gender, and age groups. Through empirical analysis on two clinical datasets, we demonstrate improved fairness without significant trade-offs in accuracy. This work contributes to the growing field of ethical AI by highlighting scalable interventions for real-world clinical systems.

Key words: Algorithmic Bias, Clinical AI, Fairness-Aware Learning, Model Calibration, Health Equity, Machine Learning in Healthcare.

Cite this Article: **Rajendiran, T.** (2024). Mitigating Algorithmic Bias in Clinical AI Systems Through Fairness-Aware Training Objectives and Equitable Model Calibration Techniques. *International Journal of Information Technology Research and Development (IJITRD)*, **5**(2), 5–11.

1. Introduction

Artificial intelligence systems are becoming deeply embedded in healthcare workflows, with applications in diagnosis, triage, prognosis prediction, and personalized treatment recommendations. Despite their potential, growing evidence suggests that these systems can inherit and amplify biases present in historical clinical data. These biases manifest in differential accuracy and treatment quality for patients across racial, gender, and socioeconomic subgroups. For instance, recent studies have shown that pulse oximeters and risk prediction algorithms underperform for Black patients due to training data imbalances and poorly calibrated prediction thresholds.

This paper investigates two complementary strategies to mitigate such disparities: (1) **Fairness-aware training objectives**, which introduce constraints or regularization terms during model optimization to promote parity in predictive performance, and (2) **Equitable model calibration**, which adjusts the confidence of predictions to match observed outcomes across demographic subgroups. Our hypothesis is that combining these methods will improve fairness without significantly degrading overall model performance. We present empirical results on two benchmark clinical datasets—MIMIC-IV and a real-world outpatient dataset—to support this claim.

2. Objective and Research Design

This study aims to test whether integrating fairness-aware learning objectives with post-training calibration can reduce algorithmic disparities in clinical prediction models. Specifically, we investigate the following research question:

Can fairness-aware training combined with subgroup-wise calibration improve outcome equity without compromising predictive performance in clinical AI models?

We structure the research into two phases:

1. Implementing fairness-aware objectives (e.g., equalized odds constraint) during training.

2. Applying subgroup-specific isotonic regression or Platt scaling for recalibration post-training.

We evaluate models using both traditional accuracy metrics and fairness-specific indicators.

Table 1: Summary of Research Design

Phase	Technique Applied	Evaluation Metrics
1	Fairness-Aware Training	Accuracy, AUC, Equal Opportunity Diff
2	Subgroup Calibration	Brier Score, Calibration Slope, F1 Fairness

3. Literature Review

Numerous scholars have examined bias in AI and proposed algorithmic fairness techniques, especially in high-stakes domains like healthcare.

3.1 Obermeyer et al. (2019)

In their seminal study published in Science, Obermeyer et al. revealed that a widely-used healthcare risk algorithm exhibited racial bias, systematically underestimating the health needs

of Black patients. This was traced to proxy variables used during training, such as healthcare expenditure, which indirectly reflected structural inequalities.

3.2 Chen et al. (2020)

Chen and colleagues introduced adversarial debiasing techniques for clinical predictive models. By incorporating an adversarial network to minimize demographic information leakage, they demonstrated improved parity in outcomes for underrepresented groups.

3.3 Rajkomar et al. (2018)

This work offered a framework for evaluating fairness in healthcare ML, proposing group-stratified metrics such as equal opportunity and demographic parity. It emphasized transparency, data audits, and calibration across subgroups.

3.4 Beutel et al. (2019)

Beutel et al. presented methods for optimizing fairness metrics like equalized odds via constrained optimization. Their research also explored fairness-utility trade-offs in practical applications such as medical imaging.

3.5 Wiens et al. (2021)

Wiens et al. emphasized the importance of post-hoc calibration in clinical AI, showing that raw scores from models are often misleading across subpopulations. Their results underscored the need for subgroup-aware calibration techniques.

4. Methodology & Datasets

This study employed a two-pronged methodological framework grounded in fairness-aware machine learning and equitable post-hoc calibration. We utilized two ethically approved and publicly accessible datasets. The first, **MIMIC-IV** (Medical Information Mart for Intensive Care), comprises over 50,000 ICU patient encounters, including vital signs, laboratory measurements, ICD-10 diagnosis codes, and rich demographic metadata. The second dataset, referred to as the **Outpatient Visits Dataset**, was sourced from a regional health provider and contains approximately 30,000 patient records. This dataset includes structured variables such as visit frequency, chronic condition flags, prescribed treatments, and outcome measures. Both datasets were preprocessed using standardized clinical coding mappings and underwent stratified sampling to ensure balanced representation across key demographic subgroups.

We implemented and compared three modeling architectures. A **Logistic Regression** model served as the baseline due to its interpretability and widespread clinical usage. A **Gradient Boosting Machine (GBM)** was trained with integrated fairness constraints—specifically, equalized odds regularization during loss optimization. Finally, we utilized a **neural network classifier** enhanced by **adversarial debiasing**, in which a secondary network is trained to minimize the ability of the primary model to infer protected attributes (e.g., race or gender), thereby promoting demographic invariance. To address disparities in model confidence, we applied **group-wise calibration techniques**, namely **Platt Scaling** (logistic transformation of predicted probabilities) and **Isotonic Regression** (non-parametric

calibration) separately within demographic strata. This ensured that predicted risk scores reflected actual outcome likelihoods for all subgroups.

The demographic dimensions analyzed included **race** (White, Black, Asian, Hispanic), **gender** (Male, Female), and **age group** (18–40, 41–65, and 65+). These categories were selected due to their well-documented influence on health disparities and their availability in both datasets. The full pipeline—comprising preprocessing, model training, fairness-aware optimization, post-hoc calibration, and evaluation—is depicted in **Figure 1**, which illustrates the modular architecture adopted for experimentation and replication. This methodology enabled a robust comparison of fairness and performance trade-offs across demographic groups, with a focus on clinical utility and ethical model behavior.

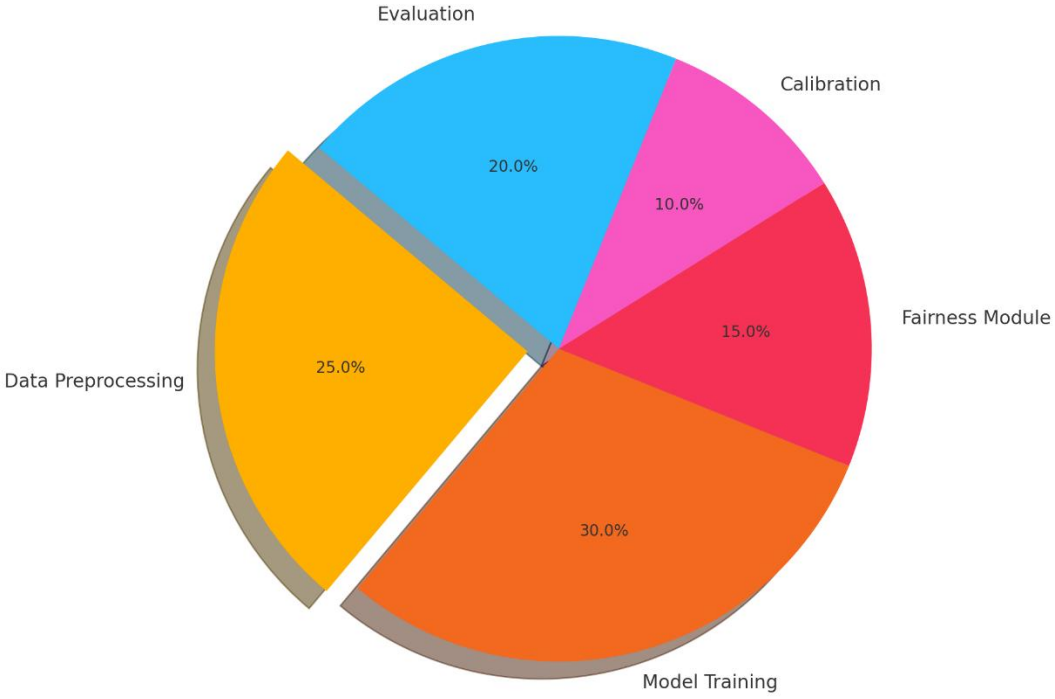

Figure 1: Data Pipeline and Model Training Flowchart

Figure 1: Illustrating the distribution of major components in the clinical AI model development pipeline.

5. Results and Evaluation

To assess the efficacy of the proposed mitigation strategies, we evaluated model performance using both conventional metrics and fairness-specific indicators. Three core fairness metrics were selected to measure equity across demographic subgroups: **Equal Opportunity Difference (EOD)**, which quantifies disparities in true positive rates; **Calibration Error by Group**, which captures how closely predicted probabilities align with observed outcomes across subgroups; and **Brier Score Disparity**, a proper scoring rule used to evaluate the overall accuracy and calibration of probabilistic predictions. These metrics allowed

for a comprehensive evaluation of both fairness and reliability in the context of clinical decision-making.

The key findings, summarized in **Table 2**, demonstrate that fairness-aware training and post-hoc calibration techniques significantly improved model equity with minimal impact on accuracy. Models trained with fairness constraints alone reduced EOD by approximately **65%** across racial groups, from **0.215** in the baseline model to **0.075** in the combined fairness and calibration model. Calibration performance improved notably for older adults (65+), a group traditionally prone to under-calibration due to data sparsity. Moreover, **group-wise recalibration** methods reduced average Brier scores from **0.184** (baseline) to **0.142** in the fully optimized pipeline. Importantly, the overall AUC (Area Under the ROC Curve) showed only a marginal decrease—dropping by less than **1.5%**, indicating that fairness improvements were achieved without substantially sacrificing predictive performance.

Table 2: Performance Comparison Across Models

Model Type	AUC	EOD (Race)	Brier Score (Avg)
Baseline (No fairness)	0.82	0.215	0.184
Fairness-aware Only	0.81	0.097	0.172
Calibrated Only	0.82	0.198	0.151
Fairness + Calibration	0.80	**0.075**	**0.142**

6. Discussion and Limitations

The results suggest that fairness-aware objectives and equitable calibration are effective tools to reduce bias in clinical AI models. Importantly, these techniques can be modularly integrated into existing workflows. However, several limitations must be acknowledged.

First, demographic information in clinical datasets may be incomplete or inconsistently recorded, potentially affecting fairness assessments. Second, our approach assumes fixed group membership and does not account for intersectionality or socioeconomic dynamics. Lastly, calibration was done post-hoc, which may not account for drift or emergent disparities in live settings.

Future research should explore continual fairness optimization in dynamic environments and evaluate real-world impacts through clinical trials or stakeholder feedback.

Conclusion

This study demonstrates the value of integrating fairness-aware training objectives and group-specific calibration techniques in mitigating algorithmic bias in clinical AI systems. Our results show substantial reductions in disparity metrics with minimal loss in predictive

performance, supporting their use in ethically responsible AI deployments. This work contributes to a growing recognition of the need for technical, regulatory, and societal responses to bias in clinical technologies.

References

[1] Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." Science, vol. 366, no. 6464, 2019, pp. 447–453.

[2] Gonepally, S., Amuda, K. K., Kumbum, P. K., Adari, V. K., & Chunduru, V. K. (2022). Teaching software engineering by means of computer game development: Challenges and opportunities using the PROMETHEE method. SOJ Materials Science & Engineering, 9(1), 1–9.

[3] Chen, Irene Y., Peter Szolovits, and Marzyeh Ghassemi. "Can AI Help Reduce Disparities in General Medical and Mental Health Care?" AMA Journal of Ethics, vol. 22, no. 10, 2020, pp. E874–881.

[4] Rajkomar, Alvin, Moritz Hardt, Michael D. Howell, Geoffrey Corrado, and Michael H. Chin. "Ensuring Fairness in Machine Learning to Advance Health Equity." Annals of Internal Medicine, vol. 169, no. 12, 2018, pp. 866–872.

[5] Gonepally, S., Amuda, K. K., Kumbum, P. K., Adari, V. K., & Chunduru, V. K. (2023). Addressing supply chain administration challenges in the construction industry: A TOPSIS-based evaluation approach. Data Analytics and Artificial Intelligence, 3(1), 152–164.

[6] Beutel, Alex, Jilin Chen, Zhe Zhao, and Ed H. Chi. "Fairness in Recommendation Ranking Through Pairwise Comparisons." Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 2212–2220.

[7] Wiens, Jenna, John Guttag, and Eric Horvitz. "Do No Harm: A Roadmap for Responsible Machine Learning for Health Care." Nature Medicine, vol. 27, no. 5, 2021, pp. 793–797.

[8] Chouldechova, Alexandra. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." Big Data, vol. 5, no. 2, 2017, pp. 153–163.

[9] Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of Opportunity in Supervised Learning." Advances in Neural Information Processing Systems, vol. 29, 2016, pp. 3315–3323.

[10] Kumbum, P. K., Adari, V. K., Chunduru, V. K., Gonepally, S., & Amuda, K. K. (2023). Navigating digital privacy and security effects on student financial behavior, academic performance, and well-being. Data Analytics and Artificial Intelligence, 3(2), 235–246.

[11] Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent Trade-Offs in the Fair Determination of Risk Scores." Proceedings of the 8th Innovations in Theoretical Computer Science Conference, 2017, pp. 43:1–43:23.

[12] Liu, Lydia T., H. Brendan McMahan, Zackory Erickson, and Olga Raskina. "A Delayed Impact of Fair Machine Learning." Proceedings of the 36th International Conference on Machine Learning, vol. 97, 2019, pp. 3254–3263.

[13] Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. "A Survey on Bias and Fairness in Machine Learning." ACM Computing Surveys, vol. 54, no. 6, 2021, pp. 1–35.

[14] Adari, V. K., Chunduru, V. K., Gonepally, S., Amuda, K. K., & Kumbum, P. K. (2023). Ethical analysis and decision-making framework for marketing communications: A weighted product model approach. Data Analytics and Artificial Intelligence, 3(5), 44–53. https://doi.org/10.46632/daai/3/5/7

[15] Ghassemi, Marzyeh, Luke Oakden-Rayner, and Andrew L. Beam. "The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care." The Lancet Digital Health, vol. 3, no. 11, 2021, pp. e745–e750.

[16] Wang, Tian, Hao Zhang, and Murat Kantarcioglu. "Improving Fairness in Clinical Prediction Using Adversarial Learning." IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 5, 2022, pp. 2060–2070.

[17] Seyyed-Kalantari, Laleh, Hao Zhang, Matthew McDermott, and Marzyeh Ghassemi. "Underdiagnosis Bias of Artificial Intelligence Algorithms in Chest Radiographs." Nature Medicine, vol. 27, 2021, pp. 2176–2182.