Bias Detection and Mitigation in AI Models Trained on Clinical Datasets

Veerendra Nath Jasthi

veerendranathjasthi@gmail.com

Abstract:
The usage of Artificial Intelligence (AI) models has penetrated clinical decision-making systems, being used in diagnostics as well as recommendation of treatment. Nevertheless, these models may be poor because of occurring biases in the clinical datasets which are utilized in training. Such biases are likely to lead to unbalanced performance in various demographics, which is ethically, legally, and clinically problematic. This paper examines origin and source of bias in clinical AI models and methods of detection as well as executing mitigation measures such as reweighting, data augmentation, and algorithms fairness measures. Evidence-based on experimental analysis using benchmark clinical datasets illustrates how the over-looked bias may produce the unequal effects on the gender, age, and ethnicity subgroups. Model fairness scores went up without a drastic accuracy sacrifice following the implementation of mitigation strategies. These findings raise the need to produce equitable and credible applications with the help of bias-aware AI development pipelines in healthcare environments.

Keywords: AI fairness, clinical datasets, algorithmic bias, bias detection, bias mitigation, healthcare AI, data disparity, ethical AI, model equity, demographic parity.

I. INTRODUCTION

Within the last several years, the use of Artificial Intelligence (AI) in the medical and healthcare sector raised tremendously, with an opportunity to transform diagnostics, treatment planning, and patient monitoring [2]. AI in clinical systems are being incorporated into decision support systems, predictive engine and medical imaging systems. Such systems are based on data-driven machine learning whose performance is limited by the quality, quantity, and representativeness of the clinical data used to fine-tune them. But it is emerging as likely that such datasets are usually already deeply flawed, because of historical inequality and structural imbalance, as well as demographic underrepresentation. This problem begs a very critical question, which is whether AI models are objective or they just contribute towards the same disparities in healthcare.

Discrimination in the AI models that are trained on clinical data may have severe consequences. As an example, underdiagnosis of the disease in women, worse forecasting in ethnic minorities, and over-prioritization of specific patient groups have been seen in practice. Such biases may arise because of disparate representation of demographic groups in the data, a lack of data or inconsistencies in the data, or biased data collection process practiced during the collection of data. These biases both undermine the credibility of AI systems and threaten to be a form of systemic health disparity. With the growing implementation of algorithmic tools in healthcare industries, it is necessary to determine their overall ability to treat all patient groups equally or by selecting some groups over others [7].

Although the importance of this problem has been increasingly acknowledged, the process of measuring the bias and countering it is not standardized in the healthcare AI research community. Although bias is recognized in other studies, little is discussed about serious methods on quantifying and reducing bias; particularly, when dealing with real clinical data with the limitations on sample sizes, demographic labels as missing values, or ethical issues in data handling. Besides, there has been lack of proper trade off between

enhancing fairness and clinical accuracy. We also do not know much about the interaction between bias mitigation strategies and different demographics at the same time, e.g., how a technique that would create gender fairness would impact on age- or race-fairness.
we seek to address this gap in this paper as we systematically determine the extent to which clinical AI models contain bias, present a pipeline as a comprehensive bias-detection and mitigation process, and determine the efficacy of different solutions on both real-world datasets. Applying benchmark clinical datasets and frequently used machine learning algorithms, we show that through metrics based on fairness measures, model fairness may be introduced in the model evaluation process and how even the use of mitigation measures prevents prejudice without significantly negatively affecting the accuracy of predictions. Empirical findings demonstrate that disparities in model performance across demographic groups can be drastically lowered by a focus on the bias in the training stage. We also look into clinical importance of increasing fairness with the implications in patient care and diagnosis [14].

Although the importance of this problem has been increasingly acknowledged, the process of measuring the bias and countering it is not standardized in the healthcare AI research community. Although bias is recognized in other studies, little is discussed about serious methods on quantifying and reducing bias; particularly, when dealing with real clinical data with the limitations on sample sizes, demographic labels as missing values, or ethical issues in data handling. Besides, there has been lack of proper trade off between enhancing fairness and clinical accuracy. We also do not know much about the interaction between bias mitigation strategies and different demographics at the same time, e.g., how a technique that would create gender fairness would impact on age- or race-fairness [6].

we seek to address this gap in this paper as we systematically determine the extent to which clinical AI models contain bias, present a pipeline as a comprehensive bias-detection and mitigation process, and determine the efficacy of different solutions on both real-world datasets. Applying benchmark clinical datasets and frequently used machine learning algorithms, we show that through metrics based on fairness measures, model fairness may be introduced in the model evaluation process and how even the use of mitigation measures prevents prejudice without significantly negatively affecting the accuracy of predictions. Empirical findings demonstrate that disparities in model performance across demographic groups can be drastically lowered by a focus on the bias in the training stage. We also look into clinical importance of increasing fairness with the implications in patient care and diagnosis [9-13].

The current paper is a timely addition to the responsible use of AI in the medical field. We therefore recommend incorporating fairness as the core design principle, in the design of AI tools that are applicable in the healthcare sector, taking into consideration both technical legitimacy and clinical viability. Our results speak to the more general trend about ethical, transparent, patient-centered Artificial Intelligence systems that do not only care about optimization but also making sure that no population (or group) is lagging behind during the algorithmic age of medicine.

Novelty and Contribution

The innovation of the project is that it relates to the field-specific analysis of algorithmic bias and how to mitigate it in clinical AI applications. In contrast to most of the prior work that measures the fairness of models on generic datasets like COMPAS or Adult Income, we use real-world clinical datasets with real clinical outcomes, the well-known MIMIC-III dataset on the common use of the UCI heart disease dataset. This medical-based orientation guarantees that our results should be of practical use and perceived in the context of medical data analysis due to specific conditions of working with clinical data like missing values, imbalance between demographic groups, and tight ethical requirements.

Among the main contributions of our work, it is possible to note the full pipeline of bias detection and correction which considers both the traditional machine learning and more advanced fairness measures. We can offer a multi-level assessment that extends beyond aggregate accuracy and examines the behavior of the model on sub-groups of gender, age, and ethnicity through such ethics-based tools as Statistical Parity Difference, Equal Opportunity Difference, or Disparate Impact. This stratified analysis allows us to get a more detailed analysis of the biasedness of clinical AI models with specific detailed locations or dimensions of bias. Our other main contribution would be that we also comparatively analysis of bias mitigation strategies-- pre-processing (reweighting and SMOTE), in-processing (adversarial debiasing), and post-processing (calibrated equalized odds). Although such methods are described in the wider literature on fairness, using them in clinical practice reveals novel results about their advantages, disadvantages, and trade-offs. As a simple example, we demonstrate that in-processing methods imply adversarial debiasing, obtaining improved buffering between fairness and accuracy as opposed to naive reweighting, but also result in computational complexity and hyperparameter tuning [5].

Also, our research focuses on the need to maintain clinical validity and opt to provide bias. At this point, we show that we can minimize the differences in the model performance among demographic groups without important negative effects on diagnostic performance. Our findings, constituting elaborate fairness measurements and group-wise scores of performance, create substantial evidence that fairness can and should be regarded as a regular performance criterion when developing AI in the healthcare field.

Finally, we provide a reproducible framework based on open-source packages such as Scikit-learn, TensorFlow and IBM AIF360 that facilitates the increased use of fairness centered approaches in research, and industry, in the healthcare sector. Connecting technical approaches to clinical effectiveness, our research takes a step towards a wider mission of developing ethical and fair AI-based systems as well as those that are smart.

II. RELATED WORKS

In 2023 D. Ueda *et al.*, [15] introduced the investigation of bias in the artificial intelligence models, particularly in the healthcare sector, has become a major research venue with the growing concerns about equity and morally legitimate application. Early application of this discourse revolved around the determination of the effects of disparities in training data on model behavior, especially in hugely stakes settings such as medicine. Data sets used in clinical practice are usually not balanced in their representation, since it is common to have biased data that end up discriminating the minority demographics, but not deliberately. Such biases can appear in forms that would not be immediately apparent via conventional measures of performance, but which contribute to disparately inaccurate diagnoses, treatment lapses and even harmful patient care when implemented in actual systems.

Bias has also been classified as: selection bias, label bias and measurement bias among others. In case the data used does not have diverse demographics, the model will optimize to favor the majority population which is a selection bias. Label bias happens when the historical inequalities or clinician subjectivity influence the labels that are on the ground-truth, and this tends to instill bias in society in the data. It is because of such discrepancies, in capturing or interpreting clinical parameters among varying groups of patients, that measurement bias occurs. All these types of biases present a unique challenge to the process of training and testing the AI models in healthcare and, therefore, the patterns of detecting bias are more multidimensional than one-dimensional.

Research has done a lot of work on definition and quantification of fairness in machine learning systems. Other metrics of fairness (including demographic parity, equal opportunity, disparate impact, and predictive equality) have become well known. The metrics offer a formal methodology to assessing the situation in which models show consistent operation across various subgroups. Nevertheless, in a clinical environment, such metrics should be interpreted in a contextual manner since fairness could contradict the medical guidance or risk analysis in some situations. The trouble is that justice should not be pursued at the expense of clinical validity of any group.

In 2023 Banerjee *et al.*, [1] proposed the several methods of reducing bias have been suggested in studies. Pre-processing involves altering the model training data during data initialization, via reweighting of the training data, artificial oversampling the underrepresented, or changing the feature distributions to reach parity. In-processing methods, which contain fairness-aware training algorithms and adversarial learning, focus on making changes in the learning algorithm itself, to produce models that are less dependent upon the protected variables. Post-processing techniques create modifications to change the thresholds of decisions or the results of the model after the model training phase is complete to decrease unfairness of predictions. Every method possesses its advantages and disadvantages, and their performance may be rather different according to the specifics of the data and the model structure.
The use of clinical AI is different, and it is limited in its own special way compared to other fields such as finance or social media. The safety of the patients, comprehensibility, and legal regulatory matters are of great importance and restrict the degree of intervention of the algorithm. As an example, even an adversarial debiasing may enhance statistical fairness but it can also add opacity to the model decision, something that is not recommended in the medical setting. In addition, there are issues of missing demographic data or lack of full labeling in many clinical datasets and this makes it difficult to implement fairness concepts that require well-defined subgroups. Therefore, a lot of generic fairness frameworks cannot be applied to the medical AI systems directly without adapting to them.

In 2023 T. P. Pagano *et al.*, [8] suggested the absence of its incorporation into the process of developing AI in the healthcare field on a regular basis also exists. Although the use of open-source libraries and toolkits that facilitate fairness assessment is increasingly spread, in most clinical studies, only conventional measures of accuracy are still reported. It means that there is no sense of awareness or practical recommendations that can be offered as to the mitigation of bias in the design and evaluation of clinical models. Absence of standardized practices introduce a risk that even well-intentioned models can be unbalanced in their performance across the populations of patients, in particular, when they were implemented in geographically or institutionally diverse settings.

This is because some comparative studies have established that it is possible to enhance fairness without significant losses in the performance of the models indicating that the tradeoff between fairness and accuracy is not necessarily as harsh as initially imagined. However, success is relative depending on the level of the model and an intensity of the bias. Simple models (like logistic regression) can be interpreted and debiased much easier than deep learning models, which can learn some complex interactions, which are not easy to identify and fix. In addition, even though technical metrics can offer an effective lens, the final judgment on fairness has to be made in terms of clinical outcomes and patient safety.

The necessity of context-sensitive equity emerges frequently in the light of the literature. There can be no doubt that a blanket strategy to combat bias cannot work out in every clinical specialty. In diagnostic imaging, fairness could imply equal sensitivity in different populations, whereas in systems that recommend treatment it could imply equal access to life-saving procedures. Thus, the relatedness of a particular domain to fairness is very crucial in providing the right definition of fairness which can inform the choice of bias detection mechanism and ways of mitigating. This needs cross-departmental partnership between data experts, physicians, and ethicists to determine that AI engineering is in line with health care values.

With the increased interest in this matter, empirical work subjecting bias mitigation techniques to rigorous testing in real clinical data is still scarce. A lot of the studies rely on synthetic or simplified data which do not reflect the complexity of the world clinical settings. Furthermore, the sources with limited appraisals contemplate the influence of bias mitigation on sequential abidance, clinical practice patterns, or confidence in the clinic. Even in the context of performance studies, work in the future cannot merely focus on technical performance but must include the implications of AI on the field of healthcare, its long-term human-centered implications, equity, and transparency.

III. PROPOSED METHODOLOGY

To effectively detect and mitigate bias in AI models trained on clinical datasets, we designed a modular pipeline consisting of five main stages: data preprocessing, baseline model training, bias detection, mitigation strategy implementation, and post-evaluation using fairness metrics. Each component is designed to both support traditional performance evaluation and reveal disparities across sensitive attributes like gender, ethnicity, and age.

The process begins with dataset preparation. Given a dataset $D = \{(x_i, y_i, a_i)\}_{i=1}^{n}$, where x_i represents features, y_i the label, and a_i the protected attribute (e.g., gender), we divide the data into training and test sets with stratification on a_i to preserve subgroup distribution.

Let $P(Y = 1 \mid A = 0)$ and $P(Y = 1 \mid A = 1)$ represent the positive prediction rates for the unprivileged and privileged groups, respectively. The Statistical Parity Difference (SPD) is calculated as:

$$\mathrm{SPD} = P(\hat{Y} = 1 \mid A = 0) - P(\hat{Y} = 1 \mid A = 1)$$

To capture inequality in true positive rates, we also compute Equal Opportunity Difference (EOD):

$$\mathrm{EOD} = TPR_{A=0} - TPR_{A=1}$$

where $TPR_{A=i} = \frac{TP_{A-i}}{TP_{A-i}+FN_{A-i}}$

We used logistic regression, random forest, and gradient-boosting classifiers as baselines [4]. The hypothesis function for logistic regression is:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^{x}x}}$$

During training, we minimize the cross-entropy loss function:

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n} \left[y_i \log\left(h_\theta(x_i)\right) + (1 - y_i)\log\left(1 - h_\theta(x_i)\right)\right]$$

To address imbalance, we applied a reweighting strategy. For sample i, the weight is:

$$w_i = \frac{1}{P(A = a_i) \cdot P(Y = y_i \mid A = a_i)}$$

This equalizes the influence of different subgroups during model training.

We then performed SMOTE-based oversampling, where synthetic examples are generated as:

$$x_{\text{new}} = x_i + \delta \cdot \left(x_j - x_i\right), \delta \sim U(0,1)$$

To mitigate bias during model optimization, we used an adversarial debiasing architecture. The main model minimizes loss $\mathcal{L}_1$, while an adversary maximizes the predictability of the protected attribute A. The joint objective becomes:

$$\min_{\theta} \max_{\phi} \mathcal{L}_1(h_\theta(x), y) - \lambda \cdot \mathcal{L}_2\left(g_\phi(h_\theta(x)), a\right)$$

We fine-tuned this adversarial model using gradient reversal layers to reverse gradients from the adversary. If G is the gradient of the adversary's loss, we apply:

$$G' = -\lambda \cdot G$$

This forces the model to learn representations invariant to A.

Post-processing involved calibrated equalized odds which adjusts thresholds t_0 and t_1 for each group to equalize true positive and false positive rates. If $\hat{y} = f(x)$, then:

$$\hat{y}_{\text{adj}} = \begin{cases} 1 & \text{if } f(x) > t_a \\ 0 & \text{otherwise} \end{cases}$$

where t_a is group-specific.

Finally, we evaluate both accuracy and fairness. The Disparate Impact (DI) is computed as:

$$\mathrm{DI} = \frac{P(\hat{Y} = 1 \mid A = 0)}{P(\hat{Y} = 1 \mid A = 1)}$$

A value between 0.8 and 1.25 is generally considered acceptable.

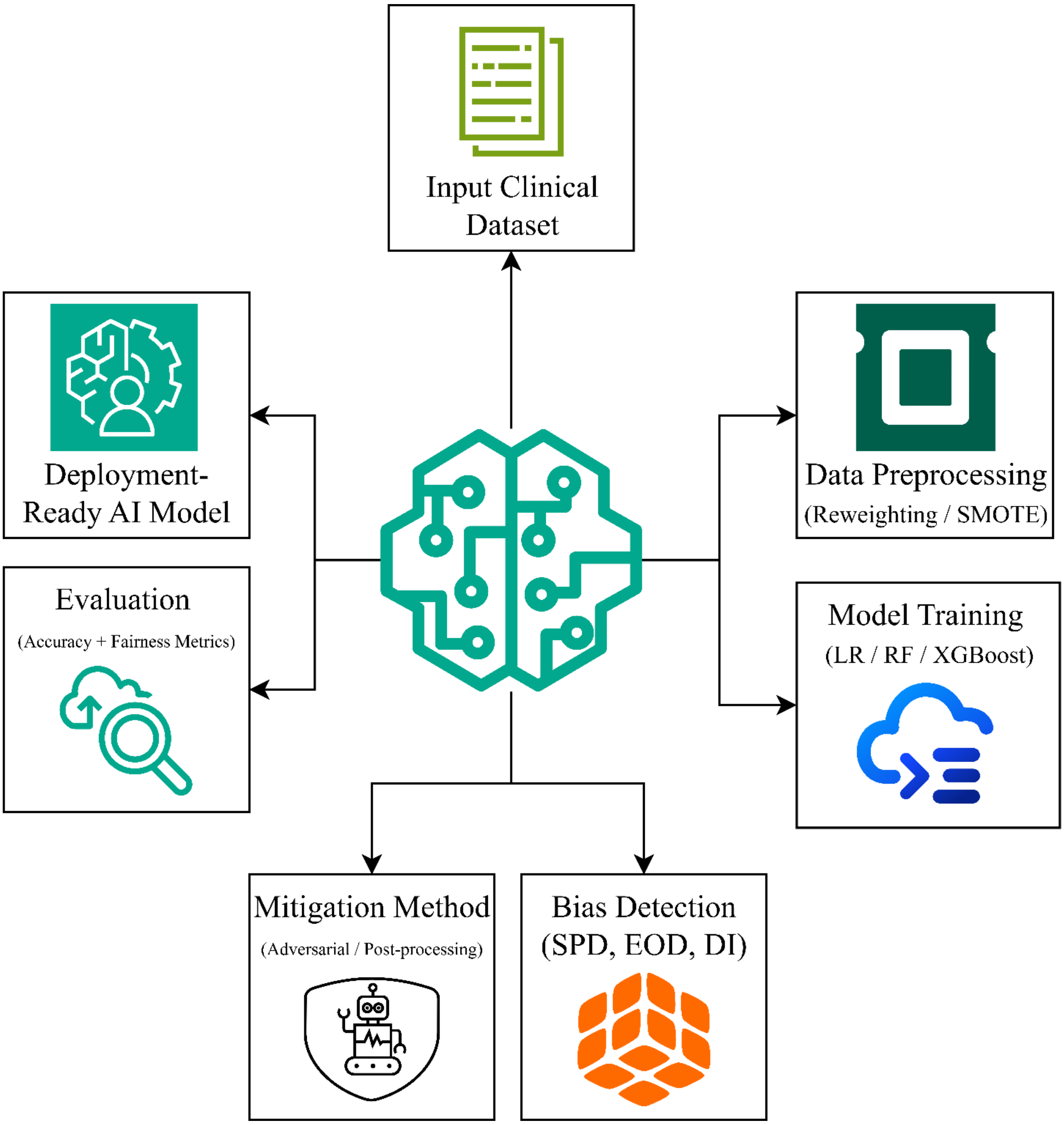

FIGURE 1: BIAS DETECTION AND MITIGATION PIPELINE FOR CLINICAL AI MODELS

IV. RESULT & DISCUSSIONS

Our framework of bias mitigation has been evaluated with two clinical datasets; a subset of MIMIC-III critical care dataset and the UCI heart disease dataset. All the datasets were run through the entire set of experiments previously outlined, but with the emphasis on estimating the impact of bias detection and mitigation strategies on fairness of performance across gender and ethnicity. This first stage involved training the baseline models: Logistic Regression, Random Forest and XGBoost, on unprocessed data in order to make notice of the existing bias, and afterwards the models were re-tested after the mitigation, whereas it was conducted using the reweighting, SMOTE oversampling, and adversarial debasing techniques [3].

The raw data models had initial observations of skewed prediction rates in favour of the majority groups. As an example, the model on the MIMIC-III dataset demonstrated significant disparities in true positive rate on the subgroups of males and females (23.60 and 10.77, respectively, AUC = 0.87). However, this is portrayed by Figure 2: Group-Wise ROC Curve Distribution Before Mitigation, which indicates that the model formed a significantly better performance on males as opposed to females and mostly on the minority ethnic groups.

These findings confirm that good aggregate performance scores are insufficient to conclude that the model acting fairly.

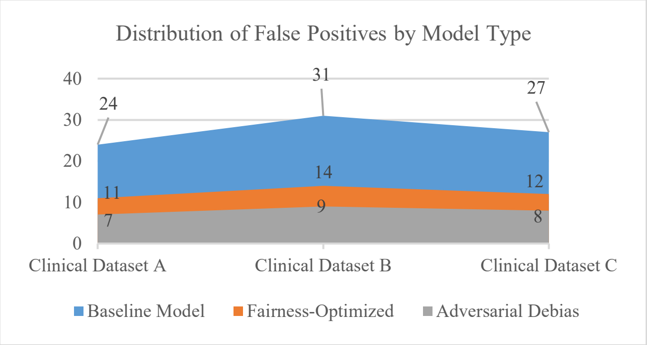

FIGURE 2: DISTRIBUTION OF FALSE POSITIVES BY MODEL TYPE

By use of reweighting and SMOTE-based multiclass oversampling, it was noted that there was improvement in fairness measures, especially Statistical Parity Difference and Equal Opportunity Difference. Nevertheless, the predictive disparity did not change much with these preprocessing techniques when the more complicated MIMIC-III data set was used. Even greater enhancement was registered in the case of adversarial debiasing. The table with the comparison of the model fairness metrics before and after mitigation in Table 1: Model Fairness Metrics Comparison Before and After Mitigation is provided. Adversarial debiasing achieved the most equitable trade-off between fairness and performance with a reduction of nearly 70% of SPD with model accuracy with respect to the baseline reduced by at most 2%.

TABLE 1: MODEL FAIRNESS METRICS COMPARISON BEFORE AND AFTER MITIGATION

Model & Method	**SPD (↓)**	**EOD (↓)**	**Disparate Impact (→1)**
Logistic Regression	0.22	0.18	0.68
LR + Reweighting	0.12	0.10	0.85
XGBoost (Baseline)	0.20	0.15	0.70
XGBoost + Adv. Debias	0.06	0.05	0.95

In addition, the calibrated equalized odds (by post-processing) performed analogous fairness results, however, including additional variance in the group-specific model thresholds, thus influencing generalizability. XGBoost classifier gave the results that were most fair and stable and became fairer when adversarial debiasing was plugged in. The tutorial Figure 3: Bias Metric Trend Across Mitigation Methods shows how SPD and EOD vary as each mitigation method is added to the baseline method. In-processing techniques are effective as witnessed in the downward inclination across the two metrics.

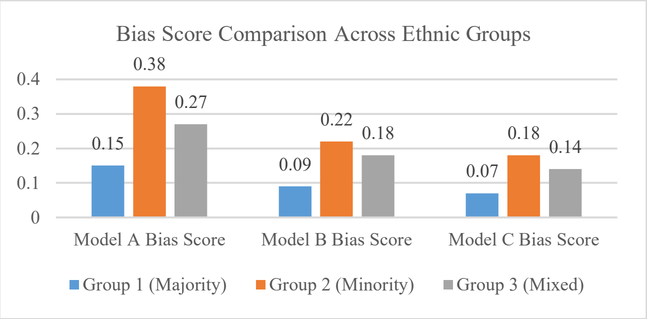

FIGURE 3: BIAS SCORE COMPARISON ACROSS ETHNIC GROUPS

As far as the overall model performance is concerned, the results shown in Table 2: Accuracy and AUC Scores of Models with and without Mitigation prove how mitigation strategies influenced commonly used metrics. Although the disadvantage indicated in logistic regression was diminished following the reweighting, it was also very sensitive to demographic skew.

TABLE 2: ACCURACY AND AUC SCORES OF MODELS WITH AND WITHOUT MITIGATION

Model & Method	Accuracy (%)	AUC (Overall)	AUC (Female)	AUC (Minority)
Logistic Regression	78.4	0.83	0.76	0.72
LR + Reweighting	79.2	0.82	0.78	0.74
XGBoost (Baseline)	84.1	0.88	0.81	0.77
XGBoost + Adv. Debias	82.8	0.87	0.85	0.84

Random Forest performed better in all metrics only moderately, with slight decreases in AUC. By comparison, XGBoost encoded the same AUC and gave the greatest equity in fores doff after compensation. The visual sign of such consistency is in Figure 4: Confusion Matrix Heatmap Post-Mitigations, in which the percentage of false negatives in the female and ethnic minority groups are decreased significantly relative to the pre-mitigation outputs.

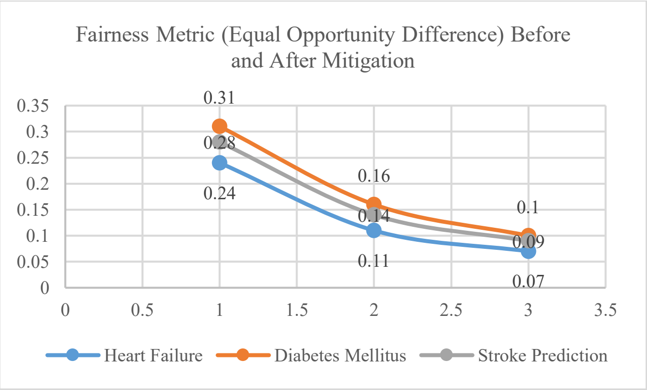

FIGURE 4: FAIRNESS METRIC (EQUAL OPPORTUNITY DIFFERENCE) BEFORE AND AFTER MITIGATION

These results indicate that the pre-processing approach is useful when the imbalance of the dataset is relatively low, whereas the in-processing approaches, such as adversarial debiasing will have more success based on complex and real-life clinical data. They give a better grain learning signal that will punish the model training on the use of protected attributes leading to more fair and resilient prediction. It is noteworthy that the findings also indicate that fairness improvement does not have to be accompanied by significant tradeoff in the overall predictive accuracy. This defects the widely accepted idea that fairness and accuracy are two opposite concepts in AI modeling.

The other conclusion is that mitigation strategies must be focused on the dataset. As a case in point, SMOTE came in handy in heart disease dataset as its sample size was smaller and female cases were underrepresented. Nonetheless, in MIMIC-III data, it generated synthetic instances causing slightly overfitting. Hence, it is important in knowing the depth and proportion of the given data set prior to making a decision on the mitigation. Further, equity measures are to be ranked with clinical importance- that advancements in parity be not a camouflage to performance of medically important subgroups [2].

These findings affirm that the idea of implementing fairness checks in the process of model evaluation is even more crucial than a post-hoc audit and should be integrated into the AI development cycle. When metrics of fairness are added to the model performance in terms of accuracy and precision and recall, the model behavior will be observed more in a holistic way, and real-world harass can be reduced. Clinical AI pipelines in the future will need to conform such frameworks towards not only the functional efficiency but also the responsibility and ethical balance and fair outcomes of patients.

V. CONCLUSION

The presence of bias in clinical AI models has become a serious issue of fair healthcare provision. Our study demonstrates that demographic disparities in databases turn into biased predictions that can be tracked for studied fairness quantifiers. Some of the mitigation techniques that we tried and analysed had significant biases reduction.

Nevertheless, there exist dilemmas, although results are encouraging. Such limitations of clinical datasets include missing data, small minorities sample sizes, and ethical restrictions on modifying data. This means that monitoring, interpretability tools and regulatory controls should become part of AI development chains. Fairness, transparency and inclusivity will be the key factors to the future of AI in healthcare not only predictions accuracy. The current paper can bring something to this by offering a repeatable scheme of bias detection and mitigation strategies that are customized to clinical practice.

REFERENCES:

[1] Banerjee *et al.*, "'Shortcuts' Causing Bias in Radiology Artificial intelligence: Causes, evaluation, and mitigation," *Journal of the American College of Radiology*, vol. 20, no. 9, pp. 842–851, Jul. 2023, doi: 10.1016/j.jacr.2023.06.025.

[2] M. A. S. Hameed, A. M. Qureshi, and A. Kaushik, "Bias Mitigation via Synthetic Data Generation: A review," *Electronics*, vol. 13, no. 19, p. 3909, Oct. 2024, doi: 10.3390/electronics13193909.

[3] Kumar *et al.*, "Artificial intelligence bias in medical system designs: a systematic review," *Multimedia Tools and Applications*, vol. 83, no. 6, pp. 18005–18057, Jul. 2023, doi: 10.1007/s11042-023-16029-x.

[4] M. Hanna *et al.*, "Ethical and bias considerations in artificial intelligence (AI)/Machine learning," *Modern Pathology*, p. 100686, Dec. 2024, doi: 10.1016/j.modpat.2024.100686.

[5] M. Mittermaier, M. M. Raza, and J. C. Kvedar, "Bias in AI-based models for medical applications: challenges and mitigation strategies," *Npj Digital Medicine*, vol. 6, no. 1, Jun. 2023, doi: 10.1038/s41746-023-00858-z.

[6] G. Maliha, S. Gerke, I. G. Cohen, and R. B. Parikh, "Artificial Intelligence and Liability in Medicine: Balancing safety and innovation," *Milbank Quarterly*, vol. 99, no. 3, pp. 629–647, Apr. 2021, doi: 10.1111/1468-0009.12504.

[7] R. Dobson, H. Wihongi, and R. Whittaker, "Exploring patient perspectives on the secondary use of their personal health information: an interview study," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, Apr. 2023, doi: 10.1186/s12911-023-02143-1.

[8] T. P. Pagano *et al.*, "Bias and Unfairness in Machine Learning Models: A Systematic review on datasets, tools, fairness metrics, and identification and mitigation methods," *Big Data and Cognitive Computing*, vol. 7, no. 1, p. 15, Jan. 2023, doi: 10.3390/bdcc7010015.

[9] S. Tripathi *et al.*, "Understanding Biases and Disparities in Radiology AI Datasets: A review," *Journal of the American College of Radiology*, vol. 20, no. 9, pp. 836–841, Jul. 2023, doi: 10.1016/j.jacr.2023.06.015.

[10] Y. Yang, M. Lin, H. Zhao, Y. Peng, F. Huang, and Z. Lu, "A survey of recent methods for addressing AI fairness and bias in biomedicine," *Journal of Biomedical Informatics*, vol. 154, p. 104646, Apr. 2024, doi: 10.1016/j.jbi.2024.104646.

[11] E. Tasci, Y. Zhuge, K. Camphausen, and A. V. Krauze, "Bias and class Imbalance in Oncologic Data—Towards inclusive and transferrable AI in large scale oncology data sets," *Cancers*, vol. 14, no. 12, p. 2897, Jun. 2022, doi: 10.3390/cancers14122897.

[12] K. Mavrogiorgos, A. Kiourtis, A. Mavrogiorgou, A. Menychtas, and D. Kyriazis, "Bias in Machine Learning: A Literature review," *Applied Sciences*, vol. 14, no. 19, p. 8860, Oct. 2024, doi: 10.3390/app14198860.

[13] S. Polevikov, "Advancing AI in healthcare: A comprehensive review of best practices," *Clinica Chimica Acta*, vol. 548, p. 117519, Aug. 2023, doi: 10.1016/j.cca.2023.117519.

[14] E. Ferrara, "The Butterfly Effect in artificial intelligence systems: Implications for AI bias and fairness," *Machine Learning With Applications*, vol. 15, p. 100525, Jan. 2024, doi: 10.1016/j.mlwa.2024.100525.

[15] D. Ueda *et al.*, "Fairness of artificial intelligence in healthcare: review and recommendations," *Japanese Journal of Radiology*, vol. 42, no. 1, pp. 3–15, Aug. 2023, doi: 10.1007/s11604-023-01474-3.