Assessing Algorithmic Bias in Machine Learning Classifiers: A Fairness Evaluation

S. Hamdan[1], S. Wagle[2], S. Poudel[2], Y. Zhou[2] and K. Poudel[1,2]

1. Computer Science, Middle Tennessee State University, Murfreesboro, TN, USA
2. Computational and Data Science, Middle Tennessee State University, Murfreesboro, TN, USA
{sah9j,sw8k,sp2ai,yz3r}@mtmail.mtsu.edu, khem.poudel@mtsu.edu

In the realm of healthcare, AI stands as an invaluable tool that significantly contributes to streamlining healthcare records [1]. It excels in organizing medical documentation, enabling remote patient monitoring, and extracting critical information from EHR systems. Furthermore, AI profoundly influences drug discovery and development by predicting patient responses to medications and identifying those who will benefit most from specific treatments [2]. This technological advancement also enhances the efficacy of personalized medicine by tailoring treatments to patients based on their medical histories.

Given the implications of AI systems in critical domains, like health, their deployment must be accompanied by careful assessments of fairness. Ensuring fairness and lack of bias in AI model predictions is crucial for gaining trust from both patients and policymakers [3]. A fair AI system should be impartial, inclusive, and transparent, considering all patient circumstances and needs. There are multiple methods that have been proposed in order to test if a system can be considered fair. One such method is known as demographic parity [4]. Demographic parity states that a model is fair if the fraction of positive predictions of a model across protected subgroups is equal. A problem with this metric is that, depending on the label, disparate outcomes could be produced. This is because, due to external factors, certain subgroups may need to be prioritized over other subgroups, and a model that follows demographic parity won't be able to capture the level of importance of a subgroup accurately. Another method is predictive parity, which states that a model is fair if the predictive positive and predictive negative values are equal between subgroups. Scores need to have observable risks between groups. If the risk distributions are not equal among groups, then it could lead to overqualification and underqualification for members in certain subgroups. Equalized odds is another metric which claims that a model is fair if its sensitivity and specificity are equalized among subgroups. Satisfying this metric can be complex if the ROC curves for subgroups don't intersect. In addition, the chances of satisfying the listed methods simultaneously is low. This means that choosing a fairness metric will be based on the dataset used.

There are multiple papers that have dealt with the challenges of fairness in healthcare. Burlina et al. dealt with AI bias in retinal diagnostics [5]. They used novel generative methods of synthetic fundus images in order to debias their AI model. Li et al. used an adversarial multi-task training strategy to successfully remove bias from their model [6]. Puyol-Anton et al. tested various methods to remove bias in cardiac MR data [7].

This paper investigates the algorithmic bias within artificial intelligence models using Diabetes 130-US Hospitals Dataset, accessed through the UCI Machine Learning Repository [8]. This dataset contains 10 years worth of clinical care data at 130 US hospitals and integrated delivery networks. The number of records is approximately 101,766, and the data has 50 features. Every record represents a patient diagnosed with diabetes whose hospital stay lasted up to two weeks. Features include demographics, length of stay, number of procedures, and more. The feature that we are targeting for fairness will be the race feature, as it is the most imbalanced. All features, excluding the encounter_id and patient_nbr, will be included when training the models. We performed three steps when preprocessing the features. Firstly, we removed missing values. For features with only a small amount of missing values, we used the mode of the feature to fill them. For features that had a significant amount of missing values, where over 90% is considered significant, the feature was removed. The next step involved converting categorical features to numerical features, which

was done by creating dummy variables. Lastly, we standardized the values using StandardScaler.

In this study, equalized odds is employed as the fairness metric. While although it may be difficult to satisfy this metric, it would ensure that subgroups are equal in both error and correctness. This is critical in healthcare, where discrepancies in false negatives, false positives, or recommendations for treatment might lead to unequal outcomes across different patient populations. To compare performance on this fairness metric, we evaluated three machines learning classifiers, including logistic regression, random forests, and XGBoost.

This paper's primary objectives are to compare different machine learning classifiers (MLCs) using the same dataset and evaluate them with identical fairness metrics.The contribution of this work lies in conducting a comparative analysis of the MLCs, especially regarding their fairness aspects, in healthcare applications. In contrast to most of the studies conducted so far, which covered only the aspects of accuracy metrics, the current study investigates MLCs on the Diabetes 130-US Hospitals dataset for fairness across racial groups using Equalized Odds. It allowed us to get insights not only into the predictive performance of the classifiers but also into their potential biases, which makes our study a valuable contribution toward equitable AI systems in healthcare. By experimenting with various popular classifiers, we aim to examine the algorithmic biases each one introduces. Figure 1 shows how we conducted our experiment.

As a preliminary result, we evaluated the fairness of three machine learning models (Random Forest, Logistic Regression, and XGBoost) across six racial groups: Caucasian, African American, Asian, Hispanic, Other, and Unknown. The XGBoost classifier consisted of 100 estimators with a max depth of 6. The LG model had a max iteration of 2000. The RF used default hyperparameters. Key fairness metrics, including Equal Opportunity, Equalized Odds, and Disparate Impact, revealed significant disparities among groups. For Random Forest, the Equal Opportunity scores ranged from 0.034 (Other) to 0.100 (Hispanic), with False Positive Rates (FPR) ranging from 20.0% (Asian) to 41.9% (Hispanic). Equalized Odds showed a True Positive Rate (TPR) as low as 0.034 (Other) and as high as 0.100 (Hispanic), while Disparate Impact revealed potential bias, particularly for the Unknown group (0.434) and Other group (0.659). In Logistic Regression, Equal Opportunity scores were lower, from 0.017 (Other) to 0.050 (Hispanic), with FPRs up to 38.7% (Hispanic). Equalized Odds showed a TPR range of 0.017 (Other) to 0.050 (Hispanic), while Disparate Impact scores indicated bias for groups such as Unknown (0.407) and Other (0.683). For XGBoost, Equal Opportunity scores ranged from 0.042 (Other) to 0.094 (Hispanic), and FPRs ranged from 20.0% (Asian) to 43.3% (Caucasian). Equalized Odds revealed TPRs from 0.042 (Other) to 0.094 (Hispanic), with Disparate Impact scores still indicating bias for Unknown (0.436) and Other (0.590). In terms of accuracy, the XGBoost model performed the best, with the RF model being second, and the LG model being last. In terms of equalized odds, the opposite is the case. The LG model performs the best overall, while the RF model performs slightly worse than the XGBoost model. Given these results, there is a tradeoff between accuracy and fairness.

More specifically, this research will provide additional detail to already published papers that use the 130-Hospitals dataset for classification by showing which models are more likely to provide unbiased results. This way, the different methods used can not only be tested based on accuracy but also fairness. These findings will contribute to the development of fairer AI systems and provide essential guidelines for the ethical deployment of AI technologies by practitioners and researchers. Therefore, this research underscores the importance of understanding and addressing algorithmic bias to achieve equitable outcomes in AI-driven decision-making processes. Future work will focus on refining these classifiers and exploring new strategies to further minimize bias, thereby enhancing the fairness and reliability of AI applications.

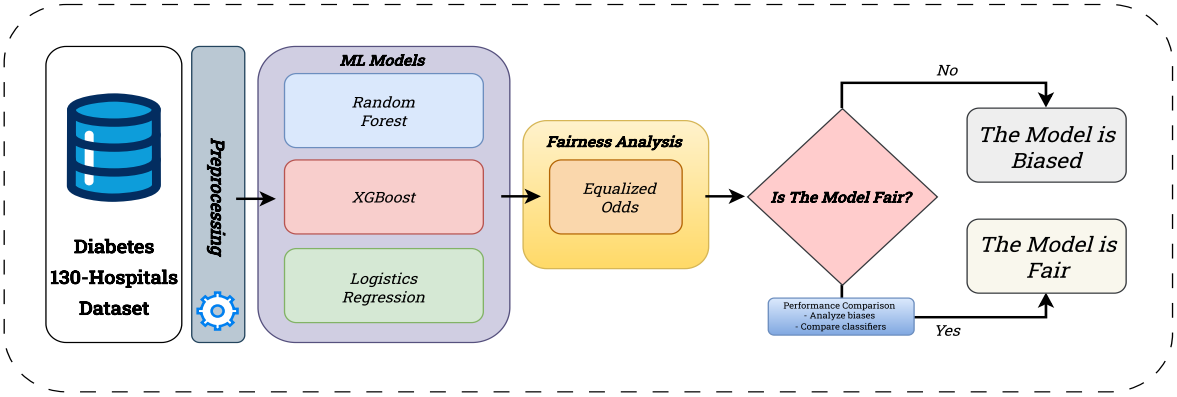

Figure 1. Block Diagram for Testing AI Models for Fairness with Fairlearn

ACKNOWLEDGEMENTS

This work was supported by Department of Computational & Data Science, and Computer Science at the Middle Tennessee State University, Murfreesboro, TN, USA.

REFERENCES

[1] A. Vardhan, "Strategic Benefits of Machine Learning Applications in Healthcare," Tezo, Jan. 10, 2024. *https://tezo.com/blog/strategic-benefits-of-machine-learning-applications-in-healthcare/* (accessed Aug. 01, 2024).

[2] "Machine Learning for Healthcare: Benefits, Use Cases & Trends," www.turing.com. *https://www.turing.com/resources/machine-learning-for-healthcare#understanding-machine-learning-in-healthcare* (accessed Aug. 01, 2024).

[3] A.-F. Näher, I. Krumpal, E.-M. Antão, E. Ong, M. Rojo, F. Kaggwa, F. Balzer, L. A. Celi, K. Braune, L. H. Wieler, and L. Agha-Mir-Salim, "Measuring fairness preferences is important for artificial intelligence in health care," The Lancet Digital Health, vol. 6, no. 5, pp. e302-e304, May 2024. [Online]. Available: *https://www.thelancet.com/digital-health*

[4] R. J. Chen et al., "Algorithmic fairness in artificial intelligence for medicine and healthcare," Nat Biomed Eng, vol. 7, no. 6, pp. 719–742, Jun. 2023, doi: 10.1038/s41551-023-01056-8.

[5] P. Burlina, N. Joshi, W. Paul, K. D. Pacheco, and N. M. Bressler, "Addressing Artificial Intelligence Bias in Retinal Diagnostics," Trans. Vis. Sci. Tech., vol. 10, no. 2, p. 13, Feb. 2021, doi: 10.1167/tvst.10.2.13.

[6] X. Li, Z. Cui, Y. Wu, L. Gu, and T. Harada, "Estimating and Improving Fairness with Adversarial Learning," 2021, arXiv. doi: 10.48550/ARXIV.2103.04243.

[7] E. Puyol-Anton et al., "Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation," 2021, arXiv. doi: 10.48550/ARXIV.2106.12387.

[8] J. Clore, K. Cios, J. DeShazo, and B. Strack. "Diabetes 130-US Hospitals for Years 1999-2008," UCI Machine Learning Repository, 2014. [Online]. Available: https://doi.org/10.24432/C5230J.

Assessing Algorithmic Bias in Machine Learning Classifiers: A Fairness Evaluation

[1] Department of Computational and Data Science Murfreesboro TN, 37132, USA;
[2] Department of Computer Science, Murfreesboro TN, 37132, USA;

S. Hamdan, S. Wagle, S. Poudel, Y. Zhou and K. Poudel

1. Introduction

In the realm of healthcare, AI stands as an invaluable tool that significantly contributes to streamlining healthcare records [1]. It excels in organizing medical documentation, enabling remote patient monitoring, and extracting critical information from EHR systems. Furthermore, AI profoundly influences drug discovery and development by predicting patient responses to medications and identifying those who will benefit most from specific treatments [2]. This technological advancement also enhances the efficacy of personalized medicine by tailoring treatments to patients based on their medical histories.

2. Data and Methods

- This paper investigates the algorithmic bias within artificial intelligence models using the Diabetes 130-US Hospitals Dataset, accessed through the UCI Machine Learning Repository[3]
- This dataset contains 10 years of clinical care data at 130 US hospitals and integrated delivery networks.
- The number of records is approximately 101,766, and the data has 50 features.
- very record represents a patient diagnosed with diabetes whose hospital stay lasted up to two weeks.
- Features include demographics, length of stay, number of procedures, and more
- The feature that we are targeting for fairness will be the race feature, as it is the most imbalanced
- We performed three steps when preprocessing the features.
- Firstly, we removed missing values. For features with only a small amount of missing values, we used the mode of the feature to fill them.
- We standardized the values using Standard Scaler.
- By experimenting with various popular classifiers, we aim to examine the algorithmic biases each one introduces. Figure 1 shows how we conducted our experiment.

3. Data and Methods (contd..)

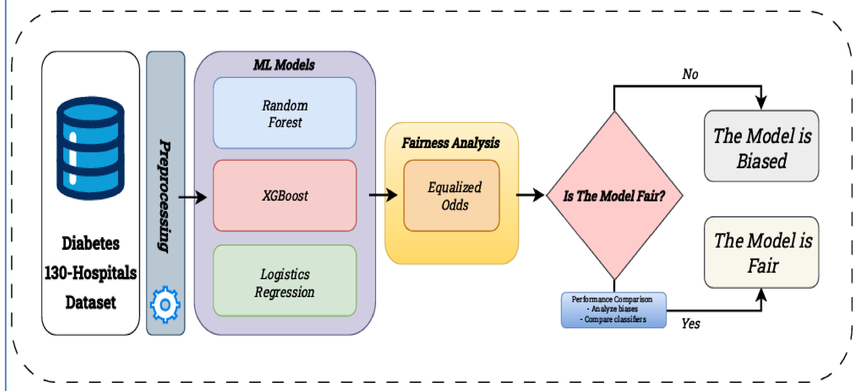

Figure 1. Block Diagram for Testing AI Models for Fairness with Fair learn.

- We evaluated three machine learning classifiers, including logistic regression, random forests, and XGBoost.
- This paper's primary objectives are to compare different machine learning classifiers (MLCs) using the same dataset and evaluate them with identical fairness metric

4. Preliminary Results

- We evaluated the fairness of three machine learning models (Random Forest, Logistic Regression, and XGBoost) across six racial groups: Caucasian, African American, Asian, Hispanic, Other, and Unknown.
- The XGBoost classifier consisted of 100 estimators with a max depth of 6. The LG model had a max iteration of 2000.
- In Logistic Regression, Equal Opportunity scores were lower, from 0.017 (Other) to 0.050 (Hispanic), with FPRs up to 38.7% (Hispanic).
- In terms of accuracy, the XGBoost model performed the best, with the RF model being second and the LG model being last.

4. Preliminary Results (contd..)

- The LG model performs the best overall, while the RF model performs slightly worse than the XGBoost model.
- This research will provide additional detail to already published papers that use the 130-Hospitals dataset for classification by showing which models are more likely to provide unbiased results.

5. Conclusion and Future work

- This project will show that different methods used can not only be tested based on accuracy but also fairness.
- These findings will contribute to the development of fairer AI systems and provide essential guidelines for the ethical deployment of AI technologies by practitioners and researchers.
- In future work, we will focus on refining these classifiers and exploring new strategies to further minimize bias, thereby enhancing the fairness and reliability of AI applications.

Acknowledgements

This work was supported by the Department of Computational and Data Science as well as the Department of Computer Science.

References

[1] A. Vardhan, "Strategic Benefits of Machine Learning Applications in Healthcare," Tezo, Jan. 10,2024. https://tezo.com/ blog/ strategic-benefits-of-machine-learning-applications-in-healthcare/(accessed Aug. 01, 2024).
[2] Machine Learning for Healthcare: Benefits, Use Cases &Trends,"www.turing.com.https://www.turing.com/resources/machine-learning-for-healthcare#understandingmachine-learning-in-healthcare (accessed Aug. 01, 2024)
[3] J. Clore, K. Cios, J. DeShazo, and B. Strack. "Diabetes 130-US Hospitals for Years 1999-2008,"UCI Machine Learning Repository, 2014. [Online]. Available: https://doi.org/10.24432/C5230J
[4] E. Puyol-Anton et al., "Fairness in Cardiac MR Image Analysis: An Investigation ofBias Due to Data Imbalance in Deep Learning Based Segmentation," 2021, arXiv. doi:10.48550/ARXIV.2106.12387