Mitigating Bias in Ischemic Heart Disease Classification Using AIF360

Jona Fejzaj
December 4, 2024

---

1. Introduction

As artificial intelligence (AI) becomes increasingly ubiquitous in our society, healthcare is undergoing significant transformation with the growth of machine learning. Models have been developed to analyze X-ray images, aid in diagnosis, and assist physicians in developing plans of care. Many of these advancements are revolutionizing healthcare as we know it, however, bias in these models can continue to perpetuate systemic inequalities in healthcare. Such cases, like the ImpactPro algorithm used to recommend patients for special care programs, highlight important lessons for researchers regarding bias mitigation and model fairness [1].

Due to the varying and unique characteristics of patients, it is crucial to mitigate biases in models that assist physicians, ensuring patients receive equitable care regardless of race or gender. Datasets are the foundation of a machine learning model's predictive capabilities and are often small or limited, especially in healthcare settings where a patient's privacy is crucial. Data constraints include lack of patient and geographical diversity due to inadequate data-sharing regulations among researchers and model developers [2]. Root causes of underrepresented groups in datasets include impediments to accessing healthcare, refusal of permission to release data, and lack of medical records [2].

One area of healthcare that can benefit significantly from machine learning is the prevention and treatment of cardiovascular diseases, which are among the leading causes of death worldwide [3]. Ischemic heart disease, also known as coronary heart disease, caused over 9 million deaths in 2019 alone [3]. Machine learning algorithms for early prediction hold great potential in identifying individuals at risk, allowing healthcare professionals to intervene with a more timely plan of care.

AI Fairness 360 (AIF360), developed by IBM, is an open source library used to determine and mitigate bias during various stages of model development [4]. This toolkit provides functionality to calculate fairness metrics and includes various pre-training and post-training algorithms designed to increase fairness and minimize bias. Sample reweighing, a pre-training technique, has shown promising results in improving fairness metrics. Tools like AIF360 can be leveraged by researchers to mitigate bias when training models with sensitive data, especially in healthcare settings. We aim to leverage AIF360's reweighing algorithm to evaluate its impact on fairness metrics while ensuring the robustness of our Ischemic heart disease classification models based on methodology from Blow et al. [5].

2. Methodology

2.1. Dataset

We used the heart disease dataset, curated in 1988, containing data from the United States, Hungary, and Switzerland [6]. Our model has been trained on the Cleveland subset of the data which contains 1025 entries, 713 of which are male (1) and 312 of which are female (0). The *'target'* feature contains heart disease diagnosis. A value of 0 represents < 50% diameter narrowing, indicating no significant heart disease, while a value of 1 represents > 50% diameter narrowing, indicating the presence of heart disease.

2.2. Random Forest

Random forest (RF) is a classification algorithm that utilizes numerous randomized decision trees to make predictions [7]. We trained two RF models, one before and one after reweighing. Grid search was utilized for hyperparameter fine tuning to optimize the models' hyperparameters. The parameter grid featured the number of estimators, max depth, and max features. Each model was trained on 80% of the data and tested on 20%.

The models' performance was evaluated using accuracy, precision, and recall. Since the primary goal of these models is to predict a diagnosis, achieving high recall is crucial to minimize the risk of misdiagnosing a patient with heart disease.

2.3. Fairness Metrics

Statistical parity difference (SPD) and disparate impact (DI) were calculated to assess the fairness of the dataset before and after reweighing. Average odds difference (AOD), Theil index (TI), and equal opportunity difference (EOD) were calculated after both models were trained.

SPD (eq. 1) measures the difference in favorable outcomes between unprivileged and privileged groups [5]. A value greater than 0 indicates more favorable outcomes for the privileged group.

$$Pr(Y = 1|D = unprivileged) - Pr(Y = 1|D = privileged) \quad (1)$$

DI (eq. 2) represents the ratio of favorable outcomes in the unprivileged and privileged group [5]. A value of 1 indicates equal outcomes. A value greater than 1 signifies bias towards more favorable outcomes for the unprivileged group.

$$\frac{Pr(Y=1|D=unprivileged)}{Pr(Y=1|D=privileged)} \quad (2)$$

AOD is the average of differences in false and true positive rates in privileged and unprivileged groups [5]. A score above 0 suggests bias towards the unprivileged group, while 0 indicates equal fairness among both groups.

TI measures group and individual fairness [5]. Lower TI values indicate more equitably distributed classification outcomes.

EOD is the difference in true positive rates between privileged and unprivileged groups [5]. A positive value indicates bias towards the unprivileged group, while 0 indicates equal fairness for both groups.

2.4. Reweighing

AIF360 contains a reweighing pre-processing technique used to adjust the significance of data points within the overall testing and training set [4]. This process takes into account each group and label combination, namely privileged and unprivileged groups with the favorable and unfavorable label respectively. For our purposes this is males (privileged) and females (unprivileged) with and without heart disease. The reweighing algorithm returns higher weights for the underrepresented group—females and lower weights for the overrepresented group—males. These weights were used to train our second model.

3. Results

We utilized various fairness metrics calculated before and after reweighing to determine fairness improvement and bias reduction. Prior to reweighing, SPD was negative, indicating favorable outcomes for females. Additionally, DI was poor compared to an 80-100% acceptable rate, signifying bias in favor of males. AOD was positive, denoting bias towards females. However, TI and EOD were acceptable values. After reweighing all metrics of concern improved significantly.

The first model had promising evaluation metrics prior to reweighing. These evaluation metrics also improved when training the second model on the reweighed data. Recall improved most significantly, reducing false negatives from 3 to 0.

Table 1. Fairness metrics before and after reweighing the data.

Metrics	Before Reweighing	After Reweighing
SPD	-0.3036	∼0.0
DI	0.4759	0.9999
AOD	0.0306	0.0
TI	0.0055	0.0
EOD	0.0	0.0

Table 2. Grid search optimal hyperparameters before and after reweighing the data.

Grid Search	Before Reweighing	After Reweighing
Max Depth	10	None
Max Features	log2	log2
Number of Estimators	25	50

Table 3. Evaluation metrics before and after reweighing the data.

Evaluation Metric	Before Reweighing	After Reweighing
Accuracy	0.9854	1.0
Precision	1.0	1.0
Recall	0.9709	1.0

4. Conclusions

By using AIF360 we were able to calculate a comprehensive range of fairness metrics, highlighting advantages for both privileged and unprivileged groups. These metrics provide a broader perspective on the overall fairness of our dataset. Evaluation metrics on our RF model performance after reweighing the data indicate improved recall with no false negatives, a critical outcome for diagnostic accuracy.

However, data limitations pose a significant challenge. The heart disease dataset is dated and contains an insignificant amount of data compared to the millions suffering with Ischemic heart disease globally. Moreover, the dataset consists of patient information exclusively from Cleveland, Ohio—a narrow demographic subset of both the United States and the global population. In the United States, certain minority groups exhibit disproportionately higher rates of heart disease [8]. There is no indication of racial breakdown in the heart disease dataset. It is unclear whether this dataset is representative of actual occurrences and reflective of the general population in the United States. To train fair and equitable models we need to standardize data collection practices in health datasets and maintain transparency.

Overfitting is another concern, especially in healthcare settings and with a dataset of our size. While the model's accuracy is promising, the same hyperparameters may not prove as effective with a significantly larger dataset. In future work, training multiple classification models could be helpful in determining overfitting and serve as a baseline comparison. Another approach worth further investigation is multi-label active learning, a classification technique designed to help models generalize rather than memorize patterns in the data [9]. This method enhances the model's ability to make more accurate predictions on unseen data and could be useful for addressing our problem scope.

References

[1] S. Sargent, “AI Bias in Healthcare: Using ImpactPro as a Case Study for Healthcare Practitioners’ Duties to Engage in Anti-Bias Measures,” *bioethics*, vol. 4, no. 1, pp. 112–116, 2021, doi: 10.7202/1077639ar.

[2] A. Arora *et al.*, “The value of standards for health datasets in artificial intelligence-based applications,” *Nat Med*, vol. 29, no. 11, pp. 2929–2938, Nov. 2023, doi: 10.1038/s41591-023-02608-w.

[3] C. W. Tsao *et al.*, “Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association,” *Circulation*, vol. 147, no. 8, Feb. 2023, doi: 10.1161/CIR.0000000000001123.

[4] “AI Fairness 360.” Accessed: Dec. 01, 2024. [Online]. Available: https://aif360.res.ibm.com/aif360.res.ibm.com

[5] C. H. Blow, L. Qian, C. Gibson, P. Obiomon, and X. Dong, “Comprehensive Validation on Reweighting Samples for Bias Mitigation via AIF360,” *Applied Sciences*, vol. 14, no. 9, Art. no. 9, Jan. 2024, doi: 10.3390/app14093826.

[6] W. S. Andras Janosi, “Heart Disease.” UCI Machine Learning Repository, 1989. doi: 10.24432/C52P4X.

[7] C. J. Mantas, J. G. Castellano, S. Moral-García, and J. Abellán, “A comparison of random forest based algorithms: random credal random forest versus oblique random forest,” *Soft Comput*, vol. 23, no. 21, pp. 10739–10754, Nov. 2019, doi: 10.1007/s00500-018-3628-5.

[8] J. A. Leigh, M. Alvarez, and C. J. Rodriguez, “Ethnic Minorities and Coronary Heart Disease: an Update and Future Directions,” *Curr Atheroscler Rep*, vol. 18, no. 2, p. 9, Feb. 2016, doi: 10.1007/s11883-016-0559-4.

[9] I. M. El-Hasnony, O. M. Elzeki, A. Alshehri, and H. Salem, “Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction,” *Sensors*, vol. 22, no. 3, p. 1184, Feb. 2022, doi: 10.3390/s22031184.