Fairness Analysis in AI Algorithms in Healthcare: A Study on Post-Processing Approaches

Vitor Galioti Martini[1] **and Lilian Berton**[1]

[1]Instituto de Ciência e Tecnologia – Universidade Federal de São Paulo (UNIFESP)
12247-014– São José dos Campos – SP – Brazil

{galioti.martini,lberton}@unifesp.br

Abstract. *Equity in Artificial Intelligence (AI) algorithms applied to healthcare is an ever-evolving field of study with significant implications for the quality and fairness of healthcare. This work focuses on applying data analysis to investigate biases in a healthcare dataset and examining how different post-processing techniques, which are less utilized and discussed in the literature compared to pre-processing techniques, can be employed to address these biases. We analyzed the Stroke Prediction dataset, and bias was identified and analyzed along with its correlation with the data. Subsequently, post-processing techniques were applied to reduce these biases, and the effectiveness of these techniques was analyzed. It was found that while all adopted post-processing techniques reduced biases, this came at the cost of a decrease in classification accuracy and precision. Among them, the EqOddsPostprocessing technique from the AIF360 library demonstrated the least impact on model accuracy and precision.*

1. Introduction

The application of Artificial Intelligence (AI) and Machine Learning (ML) in the healthcare sector has been widely adopted with the aim of enhancing the accuracy and efficiency of diagnoses, optimizing treatments, predicting clinical outcomes, and monitoring patient progress [Esteva et al. 2019]. This advancement has piqued the interest of various organizations and companies, leading to substantial investments in this field.

The increasing integration of AI into clinical applications raises ethical concerns and questions about whether algorithms are potentially amplifying existing inequalities in the healthcare system [Chen et al. 2021]. AI has been increasingly adopted in various aspects of medical diagnosis and treatment, becoming a fundamental tool in modern medicine [Jiang et al. 2017]. However, the use of these algorithms can have negative impacts, especially on subpopulations of ethnic minorities and underrepresented communities, due to spurious data relationships or systematic biases. [Chen et al. 2021] highlights the lack of adequate regulation in the approval of AI algorithms and the absence of public policies to ensure equity in the use of algorithms in healthcare, raising the question of how to ensure that algorithms are fair and do not reinforce existing inequalities. On the other hand, the study by [Obermeyer et al. 2019] identified that private healthcare systems are using prediction algorithms that, instead of measuring disease severity, predict health costs, prioritizing those who can generate more capital for the hospital rather than those who need it most. This results in a significant racial bias, as black patients, despite having more severe health conditions, end up receiving less investment in medical care due to unequal access to healthcare. According to the authors, this problem could be mitigated by substantially increasing support for black patients, from 17.7% to 46.5%.

In this way, a new area of study called algorithmic equity or ‘fairness’ has emerged. This field aims to ensure that algorithms and machine learning models treat all individuals or groups fairly and impartially. The goal is to avoid discrimination or systematic bias in automated decision-making, striving to ensure that model predictions and outcomes are not influenced by demographic characteristics such as race, gender, age, or ethnic origin that should not be relevant to the task at hand [Rabonato and Berton 2024].

Currently, some studies indicate that bias corrections in AI algorithms have proven effective [Bellamy et al. 2018, Broder and Berton 2021], and many advancements in this area are due to the development of specific tools to address this issue. Major technology companies have developed libraries to combat biases in these algorithms. For instance, IBM created AIF360, a comprehensive library that provides tools to mitigate biases in all phases of the machine learning process. Similarly, Microsoft developed Fairlearn, a library that focuses on reducing unfair disparities in machine learning model predictions. These tools demonstrate that it is possible, and indeed worthwhile, to invest in studies and bias corrections in AI algorithms in the healthcare domain.

The goal of this work is to examine the fairness of AI algorithms, focusing on the analysis of post-processing measures, as they are less explored in the literature compared to pre and in-processing techniques, and to investigate bias occurrence in the dataset concerning the protected attribute identified by a fairness measure. Additionally, we aim to examine the correlation of this fairness measure with the other attributes of the dataset, using the SHAP library.

2. Fairness

Fairness is a principle of AI ethics that seeks to eliminate bias or discrimination concerning sensitive attributes such as gender, race, and religion, among others. This means that an algorithm is considered fair when its decisions or predictions are not unduly influenced by these sensitive attributes. Fairness is a crucial aspect to ensure that ML systems are equitable and do not perpetuate or amplify existing discrimination in society. Furthermore, fairness in ML goes beyond observable data, considering additional causal information for a better understanding and removal of discrimination [Su et al. 2022].

The problem that algorithmic fairness seeks to address is complex and has various dimensions. ML algorithms are trained on datasets that often mirror biases existing in society. This means that such algorithms may end up acquiring and replicating these biases in their decisions. Additionally, the lack of transparency and the difficulty in explaining the decision-making processes of ML models can make it challenging to identify when and how these biases are being perpetuated [Castelnovo et al. 2022, Rabonato and Berton 2024].

Therefore, algorithmic fairness is an effort to identify and mitigate these biases, ensuring that algorithmic decisions are fair and non-discriminatory. This is done through various approaches and techniques, including modifying the data used to train the algorithm (pre-processing), incorporating fairness constraints during the algorithm’s training (in-processing), and adjusting algorithmic decisions after training (post-processing). Additionally, there are three important concepts related to fairness in algorithms and automated systems: group fairness, individual fairness, and causality-based fairness, each of them having specific metrics. Group fairness aims to avoid discrimination and disparities regarding specific groups, while individual fairness seeks to treat each individual fairly, regardless of their group affiliation, and causality-based fairness focuses on analyzing causal relationships to identify and mitigate the root causes of discrimination [Castelnovo et al. 2022].

For the metrics defined in the following specifications, it is important to define the concept of the trinomial $(X, A, \hat{Y})$, where:

- X: It is a set of variables that represents the characteristics or attributes relevant to the problem under discussion. These characteristics include information about the individuals involved in the decision-making process.
- A: It is a variable that represents the attribution to a certain group or category. It can indicate characteristics such as gender, race, age, etc. This variable is used to analyze possible treatment disparities or biases towards different groups.
- $\hat{Y}$: It is a target variable or result that we want to predict or analyze. It can represent a decision, an obtained result, or an expected response.

We would like to note that fairness is a complex and even subjective concept, and there are various, often conflicting, definitions of what it means to be ”fair”. Therefore, a significant challenge in algorithmic fairness is understanding how these different definitions relate and how they can be applied in different contexts.

3. Related Work

The following studies investigate biases in machine learning algorithms applied to healthcare data. They represent a sample of the studies rather than a complete review, indicating that fairness issues occur across different datasets.

3.1. Works that explored electronic records

[Gianfrancesco et al. 2018] propose interdisciplinary approaches, continuous human engagement, and strategies to ensure diverse representation in training sets to mitigate biases. [Li et al. 2023] focus on evaluating the fairness of ML models predicting cardiovascular diseases using EHR data. Their study reveals biases favoring certain race and gender groups. To mitigate bias, they explore methods like removing protected attributes and resampling to balance training group distributions. [Pivovarov et al. 2014] identify inherent biases in laboratory testing data within EHRs, arising from uncontrolled environments and missing data patterns. They propose solutions, such as leveraging missing data patterns and conducting separate analyses for different patient groups, to enhance the understanding of patient health. [Dueñas et al. 2020] investigate biases in EHRs affecting disease diagnosis and treatment. They highlight sources of bias, including changes in diagnostic criteria over time and systematic differences between demographic groups. The solutions involve considering the order and persistence of diagnoses and using biological knowledge to test the accuracy of phenotype definitions. [Noseworthy et al. 2020] evaluate biases in an AI algorithm detecting cardiac dysfunction from ECGs. The study reveals performance disparities related to patient race, attributed to a lack of diversity in the training dataset. To mitigate bias, they assess algorithm performance across racial/ethnic subgroups and train on a homogeneous population for comparison.

3.2. Works that explored NLP

[Straw and Callison-Burch 2020] aimed to uncover biases in textual data, like electronic health notes and social media posts, analyzed using Natural Language Processing (NLP). The study revealed gender bias in medical diagnoses, with women more likely to be diagnosed with personality disorders and men with Post-Traumatic Stress Disorder (PTSD). Age and gender bias in clinical trials were also identified. To address these biases, they used bias removal techniques for NLP models, cautioning that these methods might only hide persistent biases. Fairness assessment within the model was suggested through metrics like false positives, false negatives, and statistical polarity across different datasets. [Wissel et al. 2019] investigated whether an NLP algorithm, trained on doctors' notes, produced biased recommendations for pre-surgical epilepsy evaluations. The algorithm, trained on 1,097 notes from 443 patients, showed no influence of patient race, gender, or primary language on surgical candidacy scores after adjusting for demographic and socioeconomic variables. Higher scores were associated with factors like living outside the hospital's geographic area, continuing care after 18, higher household incomes, and public insurance. The study suggested that these results likely reflected referral patterns influenced by various factors, such as the severity of the patient's condition, local specialist availability, and socioeconomic factors affecting healthcare access.

3.3. Works that explored images

[Mehta et al. 2023] investigated the fairness of machine learning models in medical image analysis, focusing on tasks like classification, segmentation, and regression. They observed biases in demographic subpopulations such as race, sex, and age and employed techniques like data balancing and distribution-robust optimization to address these issues. The fairness metric used was the ”Fairness Gap”, measuring differences in task evaluation metric values based on binary sensitive attributes. However, improvements in fairness were noted to potentially compromise uncertainty estimates associated with model predictions. [Miller et al. 2023] aimed to train AI models for diagnosing Coronary Artery Disease (CAD) from myocardial perfusion images. They identified a selection bias in the training data and addressed it by augmenting the dataset with more cases of patients without obstructive CAD, making it more representative. Model calibration was used as a fairness metric to enhance accuracy in predicting obstructive CAD in low-risk patients. [Liu et al. 2022] explored the use of deep learning in analyzing chest images for diagnosing and predicting COVID-19. The study emphasized the sources of bias in deep learning models and stressed the importance of internal and external validation for reducing bias and improving model generalizability. The lack of large-scale reference datasets was identified as a significant challenge for developing fair and unbiased models.

4. Methodology

4.1. *Dataset*

The *dataset* employed in the analysis was *Stroke Prediction* published on the website *Kaggle*[1]. The *dataset* in question is interesting for analyzing bias in AI algorithms in healthcare as it includes a variety of attributes related to the risk of stroke, one of the

[1] https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

main causes of death across the world [2]. Through it, it is possible to study how different factors - biological, behavioral, and social - can interact and contribute to the occurrence of stroke.

Additionally, some of the attributes such as gender and age, may be especially prone to bias. The prevalence of stroke increases significantly with age. Most cases occur in individuals aged 65 or older, although stroke can also affect younger people. Stroke has a slightly higher prevalence in men than in women, especially at younger ages. However, women tend to experience more fatal strokes, particularly after age 65, possibly due to their longer life expectancy.

This way, as sensitive attributes, we can consider gender, age, or both. The study by [Lisabeth et al. 2009] shows that stroke can present differently in men and women and that social and medical perceptions can influence diagnosis and treatment.

4.2. Preprocessing

Pre-processing techniques are fundamental to improving data quality and preparing it for analysis. The following preprocessing actions were performed:

- Removing null values with the *dropna* function from the panda's libraries.
- Removal of the "*Other*" gender, as there was only one record, and its removal makes it possible to transform the attribute into a binary type.
- Removal of the "id" column, which represented a unique identifier for the record.
- Transformation of categorical values into numeric values using the *fit_transform* method of the *LabelEncoder* object, which is part of the *sklearn.preprocessing* library. This is important because not all ML algorithms support categorical data.

After this treatment, it was identified the imbalance caused by the low representation of the class with AVC (*stroke* = 1). This scenario motivated the application of *Undersampling* and *Oversampling* techniques, both from the *imbalanced-learn* library.

Subsampling, reducing examples from the predominant class, reduces the imbalance presented, but can lead to the loss of important information. Therefore, to counterbalance this, oversampling using the SMOTE technique, which generates new synthetic examples of the minority class, helps to reinforce the presence of this class in the data set. This combination creates a fairer balance between the classes, potentially improving the model's ability to learn features from both, resulting in more accurate predictions [Mohammed et al. 2020].

Table 1 shows that the prevalence of positive cases was previously only 4.25%, and after applying the balancing techniques it went to 50%. This balancing enabled an improvement in the performance of the models (presented in Table 2). The focus of the study turns now to the central analysis: investigating whether the models' accuracy remains consistent across different groups and analyzing the possible presence of bias in the results.

4.3. Model selection and training

To define which of the tested algorithms would be adopted in the remainder of the study, the precision and accuracy metrics were evaluated after the pre-processing steps. XGBoost was the model chosen to continue the project, as in addition to showing the best

[2] https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

Table 1. Prevalence of positive cases in the dataset

Scenario	Prevalence
Before preprocessing techniques	0.0425
After *Undersampling*	0.0909
After *Undersampling* + *Oversampling*	0.5

Table 2. Metrics after *Undersampling* and *Oversampling*

Model	Accuracy	Precision
Decision tree (DT)	0.87	0.84
Random forest (RF)	0.93	0.90
Support Vector Machine (SVM)	0.84	0.79
Multi-layer Perceptron (MLP)	0.81	0.79
XG-Boost	0.93	0.90

results, its implementation is simple and does not require fine adjustments of hyperparameters. Although the SVM and MLP models had the potential to outperform XG-Boost in terms of metrics if the hyperparameters were optimized, the efficiency presented by XG-Boost was considered sufficient to proceed with the project objectives, as the focus was not on finding the perfect algorithm for the database, but rather investigate the presence of bias.

Once the selection was made, an algorithm was created to train and analyze the bias in the model. For this purpose, the *k-fold* cross-validation technique was used, dividing the data into 5 parts (k = 5), at each iteration the procedures were carried out:

1. Splitting training and testing data using the *split* method of the *KFold* object;
2. Data normalization and standardization with the objects *MinMaxScaler* and *StandardScaler*, respectively. This pre-processing step must stay within the *k-fold* iteration to avoid data leakage.
3. Model training with the *fit* method and predictions with the *predict* method from Scikit-learn;
4. Generation of the attribute correlation graph with the *SHAP* library;
5. Display of model accuracy and precision metrics;
6. Display of bias metrics: *Disparate Impact*, *Statistical Parity Difference* and *Equalized Odds Difference (EOD).*
7. Application of post-processing techniques;
8. Redisplay of the model's accuracy and precision metrics and bias metrics, to analyze the impact of the post-processing measure on them.

4.4. Post-processing

Three post-processing techniques were applied and evaluated individually:

- *EqOddsPostprocessing* (EOP), from the library *AIF360*: This technique focuses on equalizing probabilities, that is, ensuring equal chances between groups so that the rates of false positives and false negatives are similar between them. This method adjusts the predictions of a classifier model to align the probabilities of a positive outcome, regardless of the group. Therefore, the best way to evaluate this technique is through the EOD metric, which must be as close to zero as possible to indicate success.

- *ThresholdOptimizer*, from the *Fairlearn* library: This consists of a threshold optimization technique, the method can be configured to meet two different equity criteria: demographic parity and equal chances :
  - Demographic Parity: When configured for demographic parity, *ThresholdOptimizer* adjusts the model's decision thresholds to equalize approval rates between groups. This means that it seeks to ensure that a similar proportion of records from each group receive a positive result, regardless of the actual rate of positive results within those groups. Therefore, after applying this technique the bias metrics *Disparate Impact* should approach 1 and *Statistical Parity Difference* should approach 0.
  - Equal Chances: When the *equalized_odds* parameter is passed, *ThresholdOptimizer* adjusts the thresholds to ensure that the false positive and false negative rates are similar between the groups, as well as the *EqOddsPostprocessing* technique, from the *AIF360* library described previously.

These techniques were tested in four different scenarios:

1. Protected attribute "*gender*": In this scenario, the protected attribute was gender, with male (*gender* = 1) chosen as disadvantaged and female (*gender* = 0) as favored, because the female gender has a higher frequency than the male gender, even after balancing techniques, since these made the number of records with and without stroke equal, without taking into account the gender of the record.
2. Protected attribute "*gender*" with the addition of synthetic bias: In this scenario, in addition to replicating the scenario above in terms of the protected attribute, a synthetic bias was added to randomly change the classifier variable *stroke* from 1 to 0 in half of the male records who had a stroke. This creates an artificial disproportion in the relationship between gender and the occurrence of strokes in the database.
3. Protected attribute "*group_age*": This attribute was added to use age as a protected attribute. In this context, the *age* attribute is eliminated and replaced by *group_age*. An individual is classified with *group_age* equal to 1 if they are 60 years old or older, considered elderly, and 0 if they are younger than that. *group_age* = 0 was considered as an underprivileged group since there are considerably fewer records in these conditions with stroke compared to records with *group_age* = 1.
4. Protected attribute "*group_age*" with addition of synthetic bias: In this scenario, in addition to including the new attribute *group_age*, synthetic bias is also added to it by randomly modifying the classifier variable *stroke* from 1 to 0 in half of the records where *group_age* is 0 (younger people) and *stroke* is initially 1. This change artificially creates a lower incidence of strokes in this age group specifically.

The purpose of introducing synthetic bias into the data was to amplify existing disparities, making biases more prominent and allowing for a more thorough assessment of the post-processing methods' effectiveness. As shown in Table 3, even before the addition of synthetic bias, the distribution of stroke cases between groups was not perfectly balanced, and after the introduction of synthetic bias, the disparity in prevalence became even greater.

Table 3. Prevalence of Gender and Age Group with Synthetic Bias (WSB) and without Synthetic Bias (WOSB)

Variable	Category	WSB	WOSB
Gender	0 (female)	0.558	0.558
Gender	1 (male)	0.189	0.378
Age Group	0 (young and adults)	0.187	0.374
Age Group	1 (elderly)	0.696	0.696

4.5. Correlation analysis

Using the *SHAP* library, the correlation graph shown in Figure 1 was created. Its interpretation is made as follows:

- Vertical Axis (*features*): The model's attributes are listed on the vertical axis. Each point on a horizontal line represents an instance in the dataset.
- Horizontal Axis (*SHAP value*): The horizontal axis shows the SHAP values, which measure the impact of an attribute on the model prediction. Positive SHAP values (to the right of the vertical axis) indicate that the attribute increases the probability of class = 1 prediction, while negative SHAP values (to the left of the vertical axis) indicate a decrease in this probability.
- Colors: The dots are colored to represent the attribute value. The colors range from blue (value = 0) to red (value 1) and attributes with values between 0 and 1 are represented by the color purple.
- Point Density: The density of points along the horizontal axis represents the variation in the impact of an attribute. A dense line of dots indicates that the attribute had a similar impact across many instances, while a scatter indicates variation in impact.
- Comparison between Attributes: When looking vertically at all attributes, it is possible to compare their relative importance, where attributes listed higher represent greater importance for the model prediction than those listed lower.

Analyzing the graph, it is seen that the attribute *"age"* has the most significant impact on the model's prediction, followed by *"avg_glucose_level"* and *"BMI"*. This means that age is a highly influential factor in the model's prediction, and the distribution of points shows that higher values increase the chances of the model's prediction being positive, which was already to be expected from the prior knowledge about stroke.

Regarding the *"gender"* attribute, there is a smaller distribution of points, indicating that gender has a smaller impact on the model's prediction compared to age, for example. However, it is still visible that there is a difference in the distribution of SHAP values for different genders, suggesting that gender has some role in the model's predictions. Therefore, by considering the attributes *"age"* and *"gender"* as protected, we seek to minimize the risk that the model makes unfair or discriminatory predictions based on these intrinsically sensitive characteristics.

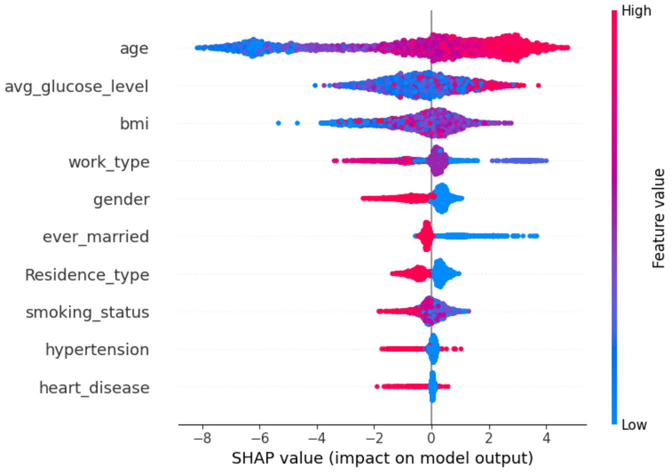

Figure 1. SHAP correlation analysis

5. Results

Tables 4, 5, 6, 7 portray the results obtained when executing the algorithm. In the columns, there are the metrics *Disparate Impact (DI)*, *Statistic Parity Different (SPD)*, *Equalized Odds Difference (EOD)*, Accuracy and Precision. In the lines, XG-Boost Classifier represents the model without applying post-processing techniques, followed by the lines that represent the techniques.

From the analysis of the results, it is clear that the application of post-processing methods in the models produces an improvement in bias metrics at the expense of accuracy and precision. This trend occurs in all tested scenarios, regardless of the presence of synthetic bias. However, in scenarios where synthetic bias is introduced, the decrease in accuracy and precision is more pronounced than in scenarios without synthetic bias. This effect is an indication that a more biased database worsens the compromise of performance metrics that post-processing methods try to correct.

For example, while XG-Boost and the Fairlearn and *AIF360* methods show a notable improvement in fairness after post-processing, accuracy and precision suffer considerable reductions, especially in the "*Age scenario group* - With Synthetic Bias" (see Table 4), where XG-Boost has an accuracy performance of approximately 0.833, which is reduced to approximately 0.685 after post-processing with the method *AIF360 EqOddsPostprocessing*. This represents a substantial drop, which is greater than that observed in the "*Gender* - With Synthetic Bias" scenario, (see Table 6), where the accuracy drops from approximately 0.879 to 0.701 with the same post-processing method. This difference may be because the age factor in the data set has a more direct impact on the prediction of the positive class, making the task of maintaining the model's performance while adjusting the bias more complex.

It is also important to highlight that between the two techniques used to balance probabilities - *EqOddsPostprocessing* from *AIF360* and *equalized_odds* from *Fairlearn* - the IBM technique (*EqOddsPostprocessing*) only failed to achieve a better Equalized Odds Difference (EOD) in the "Gender - With Synthetic Bias" scenario (Table 6). In all other scenarios, it produced results closer to zero. Nevertheless, both techniques delivered very similar accuracy and precision across all cases.

We concluded that the data analysis highlights the delicate balance between improving the fairness of models and maintaining their performance. Post-processing techniques are effective in reducing bias, but this often results in a decrease in accuracy and precision, especially in contexts where bias is more prevalent. From the data analyzed, the impact is most pronounced in the "*Age Group*" scenarios. Furthermore, the *EqOddsPostprocessing* technique from *AIF360* stood out in achieving a balance closer to the ideal in most scenarios, that is, an EOD value closer to zero, although not in all. This reinforces the idea that there is no one-size-fits-all solution for all scenarios and highlights the need for careful selection and adjustment of post-processing techniques, taking into account the specific characteristics of each dataset. Therefore, this analysis highlights the complexity involved in creating machine learning models that are both fair and effective.

Table 4. *Age group* - With Synthetic Bias, techniques Demographic Parity (DP), Equalized Odds (EO), EOPostprocessing (EOP)

Techniques	DI	SPD	EOD	Accuracy	Precision
XGBClassifier	0.189	0.602	-0.574	0.833	0.792
Fairlearn DP	0.927	0.009	0.185	0.655	0.649
Fairlearn EO	0.532	0.152	-0.015	0.690	0.666
AIF360 EOP	0.554	-0.139	-0.007	0.685	0.660

Table 5. *Age group* - No Synthetic Bias, techniques Demographic Parity (DP), Equalized Odds (EO), EOPostprocessing (EOP)

Techniques	DI	SPD	EOD	Accuracy	Precision
XGBClassifier	0.522	0.356	-0.067	0.886	0.864
Fairlearn DP	0.955	0.019	0.366	0.782	0.855
Fairlearn EO	0.714	0.203	-0.007	0.811	0.764
AIF360 EOP	0.713	-0.203	-0.005	0.811	0.764

Table 6. *Gender* - With Synthetic Bias, techniques Demographic Parity (DP), Equalized Odds (EO), EOPostprocessing (EOP)

Techniques	DI	SPD	EOD	Accuracy	Precision
XGBClassifier	0.292	0.422	-0.491	0.879	0.846
Fairlearn DP	0.941	0.034	-0.265	0.779	0.686
Fairlearn EO	0.581	0.131	-0.0008	0.706	0.775
AIF360 EOP	0.569	0.131	0.003	0.700	0.764

6. Final Considerations

This work performed a detailed analysis of the application of machine learning algorithms for a healthcare dataset. It was observed that, although each variable has its role in the classification, age proved to be the most influential factor.

Post-processing techniques, despite reducing bias, presented *trade-offs* in terms of model performance. This highlights that there is no one-size-fits-all solution for all scenarios and underlines the need for careful selection and tuning of post-processing techniques, taking into account the specific characteristics of each dataset.

Table 7. *Gender* - No Synthetic Bias, techniques Demographic Parity (DP), Equalized Odds (EO), EOPostprocessing (EOP)

Techniques	DI	SPD	EOD	Accuracy	Precision
XGBClassifier	0.680	0.190	-0.040	0.915	0.888
Fairlearn DP	0.989	0.006	-0.014	0.858	0.799
Fairlearn EO	0.755	0.139	0.004	0.891	0.872
AIF360 EOP	0.746	-0.144	-0.001	0.893	0.874

In future work, a detailed comparison can be made between pre-, in-, and post-processing techniques, focusing on their effectiveness and limitations in different contexts. Another work would be to use other databases to verify whether the behavior of the techniques remains similar in different scenarios, enabling a more robust analysis of the generalization of the methods studied. The impact on metrics when increasing records for just one group of the protected attribute could also be explored.

7. Acknowledgments

This work has been supported by Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) Grant 2021/14725-3 and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., et al. (2018). Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

Broder, R. S. and Berton, L. (2021). Performance analysis of machine learning algorithms trained on biased data. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 548–558. SBC.

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., and Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1):4209.

Chen, R. J., Chen, T. Y., Lipkova, J., Wang, J. J., Williamson, D. F., Lu, M. Y., Sahai, S., and Mahmood, F. (2021). Algorithm fairness in ai for medicine and healthcare. *arXiv preprint arXiv:2110.00603*.

Dueñas, H. R., Seah, C., Johnson, J. S., and Huckins, L. M. (2020). Implicit bias of encoded variables: frameworks for addressing structured bias in ehr–gwas data. *Human Molecular Genetics*, 29(R1):R33–R41.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., and Dean, J. (2019). A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29.

Gianfrancesco, M. A., Tamang, S., Yazdany, J., and Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, 178(11):1544–1547.

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., and Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4).

Li, F., Wu, P., Ong, H. H., Peterson, J. F., Wei, W.-Q., and Zhao, J. (2023). Evaluating and mitigating bias in machine learning models for cardiovascular disease prediction. *Journal of Biomedical Informatics*, 138:104294.

Lisabeth, L. D., Brown, D. L., Hughes, R., Majersik, J. J., and Morgenstern, L. B. (2009). Acute stroke symptoms: comparing women and men. *Stroke*, 40(6):2031–2036.

Liu, T., Siegel, E., and Shen, D. (2022). Deep learning and medical image analysis for covid-19 diagnosis and prediction. *Annual Review of Biomedical Engineering*, 24:179–201.

Mehta, R., Shui, C., and Arbel, T. (2023). Evaluating the fairness of deep learning uncertainty estimates in medical image analysis. *arXiv preprint arXiv:2303.03242*.

Miller, R. J., Singh, A., Otaki, Y., Tamarappoo, B. K., Kavanagh, P., Parekh, T., Hu, L.-H., Gransar, H., Sharir, T., Einstein, A. J., et al. (2023). Mitigating bias in deep learning for diagnosis of coronary artery disease from myocardial perfusion spect images. *European journal of nuclear medicine and molecular imaging*, 50(2):387–397.

Mohammed, R., Rawashdeh, J., and Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: overview study and experimental results. In *2020 11th international conference on information and communication systems (ICICS)*, pages 243–248. IEEE.

Noseworthy, P. A., Attia, Z. I., Brewer, L. C., Hayes, S. N., Yao, X., Kapa, S., Friedman, P. A., and Lopez-Jimenez, F. (2020). Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ecg analysis. *Circulation: Arrhythmia and Electrophysiology*, 13(3):e007988.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Pivovarov, R., Albers, D. J., Sepulveda, J. L., and Elhadad, N. (2014). Identifying and mitigating biases in ehr laboratory tests. *Journal of biomedical informatics*, 51:24–34.

Rabonato, R. T. and Berton, L. (2024). A systematic review of fairness in machine learning. *AI and Ethics*, pages 1–12.

Straw, I. and Callison-Burch, C. (2020). Artificial intelligence in mental health and the biases of language based models. *PloS one*, 15(12):e0240376.

Su, C., Yu, G., Wang, J., Yan, Z., and Cui, L. (2022). A review of causality-based fairness machine learning.

Wissel, B. D., Greiner, H. M., Glauser, T. A., Mangano, F. T., Santel, D., Pestian, J. P., Szczesniak, R. D., and Dexheimer, J. W. (2019). Investigation of bias in an epilepsy machine learning algorithm trained on physician notes. *Epilepsia*, 60(9):e93–e98.