# Analyzing the Trade-off Between Fairness and Model Performance in Supervised Learning: A Case Study in the MIMIC dataset

**Bruno Pires M. Silva, Lilian Berton**

[1]Instituto de Ciência e Tecnologia (ICT)
Universidade Federal de São Paulo (UNIFESP).
São José dos Campos – SP – Brasil

bruno.pires22@unifesp.br, lberton@unifesp.br

***Abstract.*** *Fairness has become a key area in machine learning (ML), aiming to ensure equitable outcomes across demographic groups and mitigate biases. This study examines fairness in healthcare using the MIMIC III dataset, comparing traditional and fair ML approaches in pre, in, and post-processing stages. Methods include Correlation Remover and Adversarial Learning from Fairlearn, and Equalized Odds Post-processing from AI Fairness 360. We evaluate performance (accuracy, F1-score) alongside fairness metrics (equal opportunity, equalized odds) considering different sensible attributes. Notably, Equalized Odds Post-processing improved fairness with less performance loss, highlighting the trade-off between fairness and predictive accuracy in healthcare models.*

## 1. Introduction

The application of artificial intelligence (AI) across various domains has provided numerous advantages to society [Correa et al. 2022]. However, various challenges have been reported in relation to this emerging technology. Researches have indicated that machine learning (ML) algorithms may produce biased outcomes for certain groups [Jui and Rivas 2024]. As a result, there has been an increasing focus on developing fair approaches to address these disparities and reduce the impact of interference and bias associated with demographic information in datasets.

However, one of the adverse effects of introducing methods that promote fairness in algorithmic classifications is the loss of predictive performance [Li and Li 2025], which can arise from various factors. Examples of these factors include the addition of complexity and data processing to models, the characteristics of minority groups that impact all predictions, leading to greater errors in the general case, and the mathematical conflict between different fairness metrics. These metrics may focus on individual, collective, or inter-group fairness, but are often contradictory in global terms. Thus, there is a trade-off between fairness and predictive performance in ML, which represents a significant challenge in implementing algorithms that seek fair outcomes [Barocas et al. 2018, Rabonato and Berton 2024]. Some tools have been proposed for detecting and mitigating unwanted biases in ML models, such as IBM360 [Bellamy et al. 2019] and FairLearn [Bird et al. 2020].

Medicine is an area that has applied new fairness algorithms [Taber et al. 2023], as there is a strong connection to individuals' personal characteristics and a lower tolerance

for errors. It is essential to address these issues while minimizing the impact on the overall model performance. This research utilizes the dataset MIMIC III (Multi-Parameter Intelligent Monitoring in Intensive Care) [Johnson et al. 2016], which is a valuable resource for medical research and education, it contains detailed and comprehensive data on various physiological parameters, besides demographic information such as gender, age, ethnic group, as well as laboratory test results and other relevant data. Some previous works analyzed different versions of this dataset, [Malone et al. 2018] addresses patient outcome predictions using the MIMIC-III dataset and includes pre-processing techniques to ensure fairness in the data. [Meng et al. 2022] focus on the MIMIC-IV dataset and conducts comprehensive analyses of dataset representation bias, interpretability, and prediction fairness of deep learning models for in-hospital mortality prediction. [Kakadiaris 2023] examines the fairness and bias in an XGBoost binary classification model predicting the ICU length of stay (LOS) using the MIMIC-IV dataset. [Chen et al. 2020] explores fairness in a multimodal clinical dataset, using the MIMIC-III dataset. The authors investigate fairness algorithms such as equalized odds and biased word embeddings for medical predictions.

The goal of this work is to compare various fairness techniques from IBM360 and FairLearn in disease classifications across different demographic groups from MIMIC III. The algorithms in question will incorporate techniques to maximize fairness across these groups. We aim to analyze which techniques can achieve a better balance between accurate predictions and fairness. The contributions and novelty of this work can be summarized as follows:

- Study fairness and ML with the MIMIC III dataset, which allows us to analyze different demographic groups (sex, marital status, and ethnic group ) and diseases (Congestive heart failures, Diabetic Ketoacidosis and Coronary artery disease).
- The research aims to compare various fairness techniques from IBM360 and FairLearn (Correlation remover, Adversarial learning, Equalized odds postprocessing) considering different classifiers (Gradient Boosting Classifier (GBC), Support Vector Machine (SVM), and a Multilayer Perceptron (MLP).
- The study explores which techniques can achieve a better balance between accurate predictions and fairness. This is crucial for maintaining the reliability of medical models while ensuring that they operate without bias.

## 2. Contextualization

Fairness in machine learning does not have an exact and universal definition, as what is considered fair can vary depending on the specific context. However, there are some generic definitions, concepts, and common problems that can be identified [Yeom and Tschantz 2021]. The central idea of fairness in this context revolves around ensuring equal treatment of individuals regardless of their personal characteristics. This involves examining how individuals are represented in relation to others in the analyzed data, comparing different groups, and other related factors [Ferrara et al. 2024]. Evaluation metrics are developed based on these principles to objectively identify any issues in the analyzed models or data, such as biases that may require statistical resampling techniques to address.

In machine learning models, fairness can be promoted at three stages in the pipeline. The first stage involves focusing on the data before defining the model, which

can help identify and address biases in datasets, it is called pre-processing approaches [Sattigeri et al. 2019], [Xu et al. 2019]. The second stage involves modifying existing models, such as adding new terms or adjusting the main operating logic of the algorithms to consider new parameters during the training stage, it is called in processing [Zhang et al. 2018]. The third stage involves applying fairness methods to established models, it is called post-processing.

The correlation between fairness and health has prompted the development of specific tools and models aimed at promoting equality in the healthcare sector. For instance, a racial prejudice mitigation tool was created using the operations of the Gerchberg–Saxton algorithm in the frequency domain, validated with the MIMIC-III dataset [Ay et al. 2024]. Additionally, there are studies focusing on modifications to the model definition pipeline to address bias in the clinical context [Raza 2023], along with methods targeting bias reduction in clinical images, such as electroencephalography [Kurbatskaya et al. 2023] and glaucoma detection examinations [Luo et al. 2024].

This study aims to compare the effectiveness of fairness approaches at each stage of machine learning (pre, in and post-processing) and examine the trade-off between the most beneficial points and the loss of predictive performance when applying fairness approaches, compared to the original algorithms.

## 3. Materials and methods

This study employ the MIMIC-III dataset [Johnson et al. 2016], which contains diverse information about individuals admitted to the intensive care unit at Beth Israel Deaconess Medical Center in Boston, United States. Twenty-four laboratory values obtained from patient examinations were selected to predict three diseases: diabetic ketoacidosis, heart failure, and coronary artery disease. Additionally, demographic information such as sex, marital status, and ethnic group was considered to assess variations among different groups of individuals.

The criterion for selecting the input data was primarily based on the exams with fewer null values across patients. For the classifications, the two cardiac diseases were selected due to their status as leading causes of death in Brazil [Oliveira et al. 2020], while diabetic ketoacidosis was chosen for its distinctiveness from the cardiac diseases, adding complexity to the classification while still having sufficient records in the dataset. Regarding sensitive data, the choice was made for information that allowed for good granularity, forming groups of equivalent individuals (such as the various Hispanic groups categorized as Latino), while also ensuring parsimony in the analyses and avoiding redundancy of cases.

Due to the unbalanced distribution of individuals diagnosed with each disease, it was necessary to use resampling techniques on the training dataset to ensure that the overall model results are more balanced and representative of real-world cases, although this brings risks such as potential loss of diversity, variability, and generalization of the data, higher chances of overfitting, and introduction of biases. These possible behaviors were monitored and managed to an acceptable limit for our scenario through techniques such as a combination of oversampling and undersampling and cross-validation with the real data. Figure 1 demonstrates the change in the distribution of this information, which was achieved by increasing the instances of the minority class through interpolations between

the closest points and by removing redundant occurrences, overlaps, and outliers. The algorithms used for this purpose were the Synthetic Minority Oversampling Technique (SMOTE) for promotion and Edited Nearest Neighbors for reduction (ENN).

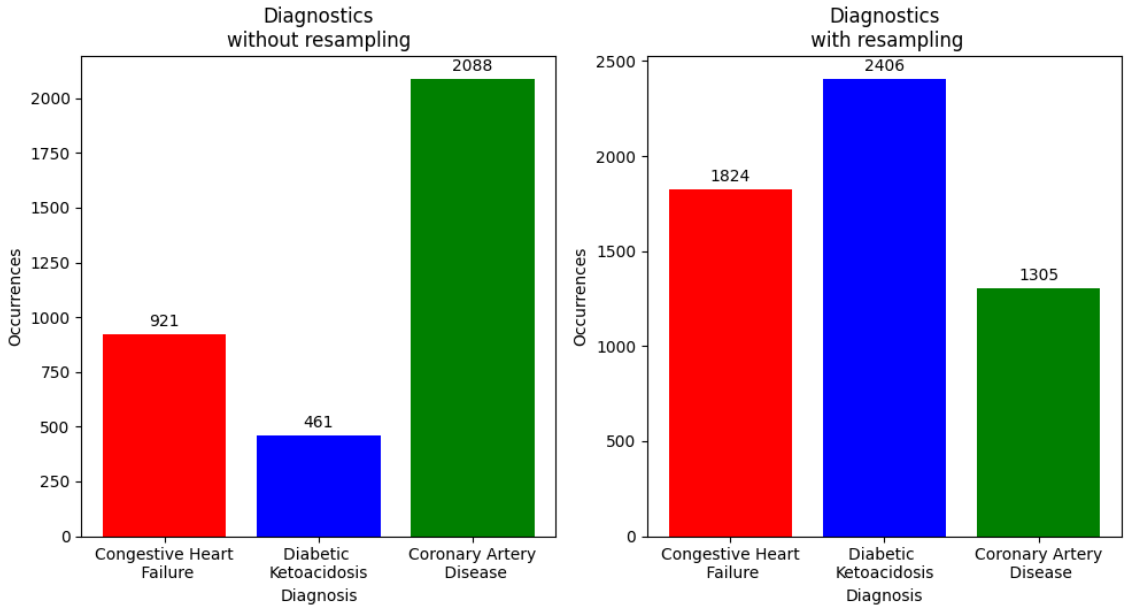


**Figure 1. Diagnosis distributions without resampling (left) and with resampling (right).**

In Figure 2, we can observe the distribution of each of the three sensitive attributes analyzed among individuals (gender, marital status and ethnic group). The data was divided with 80% for training and 20% for testing. Only the training data was resampled and used for training the machine learning models. With resampling, it is observed that the distribution of some attributes became more similar to each other; however, an imbalance still persists.

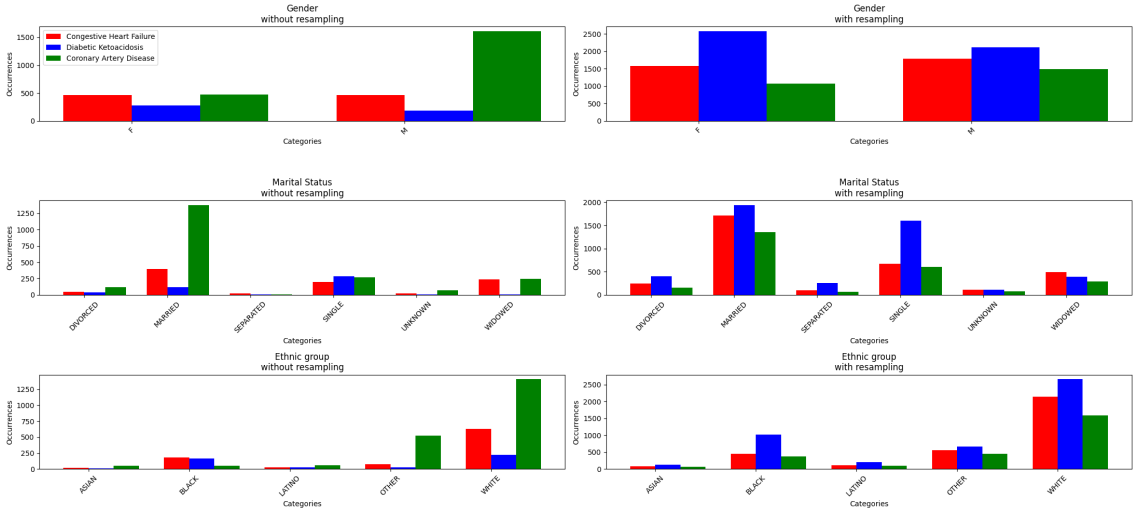


**Figure 2. Sensitive attributes distributions without resampling (left) and with resampling (right).**

The classification task used the Gradient Boosting Classifier (GBC), Support Vector Machine (SVM), and a Multilayer Perceptron (MLP). Three techniques were applied to mitigate bias, namely:

- Correlation Remover: is a technique that uses linear transformations in data sets to eliminate the correlation with the sensitive attribute. This method, which has been tested as part of the FairLearn library, is considered a data pre-processing technique because the operations are only applied to the data set before the model training phase.
- Adversarial Learning: is based on MLP networks and utilizes an adversarial approach as an in-processing method. It works by using a network to predict the main problem, and an adversarial network to determine which sensitive group the occurrence belongs to. These two networks interact by exchanging weights. The method was implemented using the FairLearn library.
- Equalized Odds postprocessing: is implemented in the AIF360 library based on linear optimization. It is used to search for probabilities of changing the output labels, find a better decision threshold, and balance the rates of true positives and false negatives among the protected groups.

To evaluate and compare the models, the traditional metrics chosen to analyze predictive performance were accuracy and F1-score. Accuracy was used to provide an overall view of the model's ability to correctly classify instances, while the F1-Score was selected due to its balance between precision and recall, which is especially relevant for dealing with imbalanced class distributions.

To assess the fairness of the models, we utilized the metrics *Equalized Odds* and *Equal Opportunity*, which measure disparities in the rates of true positives and, in the case of *Equalized Odds*, also false positives between sensitive groups. These metrics were calculated separately for the attributes of gender, marital status, and ethnicity, allowing for a detailed analysis of the impact of fairness techniques on each subgroup.

The equations for the fairness metrics are as follows:

- Equal opportunity: Based on the difference in true positive rates (TPR) between two sensitive groups, that is:

$$EqualOpportunity = \text{TPR}_{\text{Group 1}} - \text{TPR}_{\text{Group 2}}$$

With

$$TPR = \frac{TP}{TP + FN}$$

- Equalized Odds: Based on the difference in false positive rates (FPR) between two sensitive groups, that is:

$$EqualizedOdds = \text{FPR}_{\text{Group 1}} - \text{FPR}_{\text{Group 2}}$$

With

$$FPR = \frac{FP}{FP + TN}$$

## 4. Experiments

Since the MIMIC-III dataset is provided as a relational database, with various keys referencing and linking tables, these pieces of information were cross-referenced for data processing, along with handling missing values related to laboratory tests and rates. The missing values were filled using the mean of their respective columns, followed by the application of Z-Score normalization. Although this process may reduce the generalizability of the data, it helps create a more solid and robust dataset with complementary examples, even if some information is still missing, given that the dataset is not fully complete. Another preprocessing step involved merging equivalent groups, as the original dataset contained values such as white - russian and white - eastern european, which could be consolidated under the broader category white. Similarly, categories that could not be easily grouped were classified as other.

To carry out the experiments, we reformulated the predictive task as a set of binary classification problems, one for each disease. For each task, the model predicts whether a patient has a specific disease ("has disease X" vs. "does not have disease X"). This transformation follows the approach used by the fairness library and allows us to independently assess fairness and performance for each condition.

With the relevant training data properly processed, the next step was resampling, using the SMOTE and ENN algorithms. These methods were chosen due to their complementary nature, as they help preserve the unique characteristics of the data—one of the potential losses when applying such techniques. Specifically, SMOTE increases the number of minority class examples by generating synthetic samples through interpolation of nearby neighbors, based on Euclidean distance. In contrast, ENN performs the opposite task by identifying and removing instances that are considered "bad neighbors". A bad neighbor is defined as an instance that belongs to a different class than most of its surrounding neighbors, also determined using Euclidean distance. This complementary approach helps balance the dataset in a way that enhances the training process. The default parameters of the SMOTEENN implementation from Python's imbalanced-learn (imblearn) library were used, with a random state of 42. After balancing, an 80/20 train-test split with a fixed seed (42) was used, and model performance was evaluated using a hold-out strategy.

With the dataset now balanced, the next step was to evaluate different classification models under fairness constraints. A total of 21 models were created by combining three classification algorithms —Gradient Boosting Classifier (GBC), Support Vector Machine (SVM), and Multi-layer Perceptron (MLP)— with two fairness intervention stages (pre-processing using correlation remover and post-processing with equalized odds). Additionally, the MLP model incorporated an in-processing adversarial method for fairness, which was not applied to the other classifiers.

To ensure a straightforward yet effective comparison, the models were configured with simple hyperparameters that provided sufficiently satisfactory results without extensive tuning. The specific configurations were as follows:

- GBC: 100 estimators, a learning rate of 0.1, and a maximum depth of 3.
- SVM: Linear kernel with the regularization parameter $C = 1.0$ (default value).
- MLP: A single hidden layer with 20 neurons and a maximum of 500 training iterations.

## 5. Results

Due to the large number of experiments, several points could be observed from the interaction of methods, algorithms and data subsets. However, only the most relevant and common behaviors among all test variations will be presented, while all results are available on github.com/psbruno/fairness.

### 5.1. Performance variation between classifiers

When we looked at the results for each diagnosis we noticed only small differences in the traditional predictive performance, mostly in the decimal places of the accuracy percentages. This trend was consistent across all the classifiers, as shown in Table 1, for the congestive heart failure. However, there were significant variations in the fairness metrics. Not only the absolute values were different, but the scale of the variations also varied widely. Lower values were particularly sensitive to changes and discrepancies in the prediction algorithms. This sensitivity could be exploited to address errors induced by these variations.

| | **GBC** | **SVM** | **MLP** | **GBC** | **SVM** | **MLP** |
|---|---|---|---|---|---|---|
| **Phase** | Pre | Pre | Pre | Pos | Pos | Pos |
| **Accuracy** | -1.8% | -0.1% | -1.8% | -3.9% | -5.6% | -2.1% |
| **F1-Score** | -1.6% | 0% | -1.5% | -3.4% | -4.8% | -2.1% |
| **Equal Opportunity** | +3% | +2.6% | +9.4% | +3% | +3.9% | +11.4% |
| **Equalized Odds** | +2% | +8.7% | +16.7% | +6.6% | +11.5% | +11.8% |

**Table 1. Absolute variation in the performance of classifiers after applying fairness methods in the diagnosis of congestive heart failure using sex as sensitive attribute.**

It is not possible to globally define an algorithm as more efficient in terms of fairness compared to others. This is because when we compare the two metrics, we notice that they tend to conflict with each other. When there is a promotion of one particular metric, the other declines as a result. Therefore, based on the specific injustices that emerge in algorithms, we need to evaluate and prioritize which metric should be given more importance in its specific context.

### 5.2. Performance variation between diseases

Due to variations in the data sets, including issues related to the disadvantage of smaller groups, we observed differences in performance among the diagnoses. This variation is an indirect result of implementing fairness methods. While a binary approach is necessary for this purpose, it also leads to other subsequent interactions. Among all the experiments, there was a higher success rate in predicting the diagnosis of diabetic ketoacidosis. Even without the introduction of fairness steps, the average accuracy was around 97% among the algorithms. With the introduction of fairness steps, the accuracy varied between 94% and 96% for SVM, as shown in Table 2. This positive outcome is evident not only in the higher accuracies themselves but also in the smaller drop in accuracy when introducing these steps, with a maximum decrease of 3%. On the other hand, the other two diagnoses showed a greater drop in accuracy throughout the tests, ranging from 3% to 5%,

with accuracies consistently in the 80% range, significantly lower than that of diabetic ketoacidosis.

| | No fairness | Pre-Processing | Post-Processing |
|---|---|---|---|
| **Diabetic Ketoacidosis** | 97% | 96% | 93.9% |
| **Congestive Heart Failure** | 85% | 84% | 80% |
| **Coronary Artery Disease** | 88% | 87% | 83% |

**Table 2. SVM accuracy without fairness methods in different scenarios generated by the binary approach.**

### 5.3. Trade-off between fairness and predictive performance

In Figure 3, we present the results for Adversarial Learning as an in-processing approach. We observed that the performance dropped by a maximum of 10%. It was more challenging to adjust the training process to account for fairness, by introducing new terms and modifying the algorithms, especially in cases involving competition from the adversary network. Fairness metrics showed positive effects, such as a significant improvement of approximately 92.4% in equal opportunity for predicting congestive heart failure. However, this method proved to be inconsistent, as it resulted in a decrease of approximately 34.8% in equal opportunity when applied to coronary artery disease. The largest decline in accuracy was in the prediction of congestive heart failure, where the accuracy dropped from 85.7% to 74% when the in-processing method was introduced.

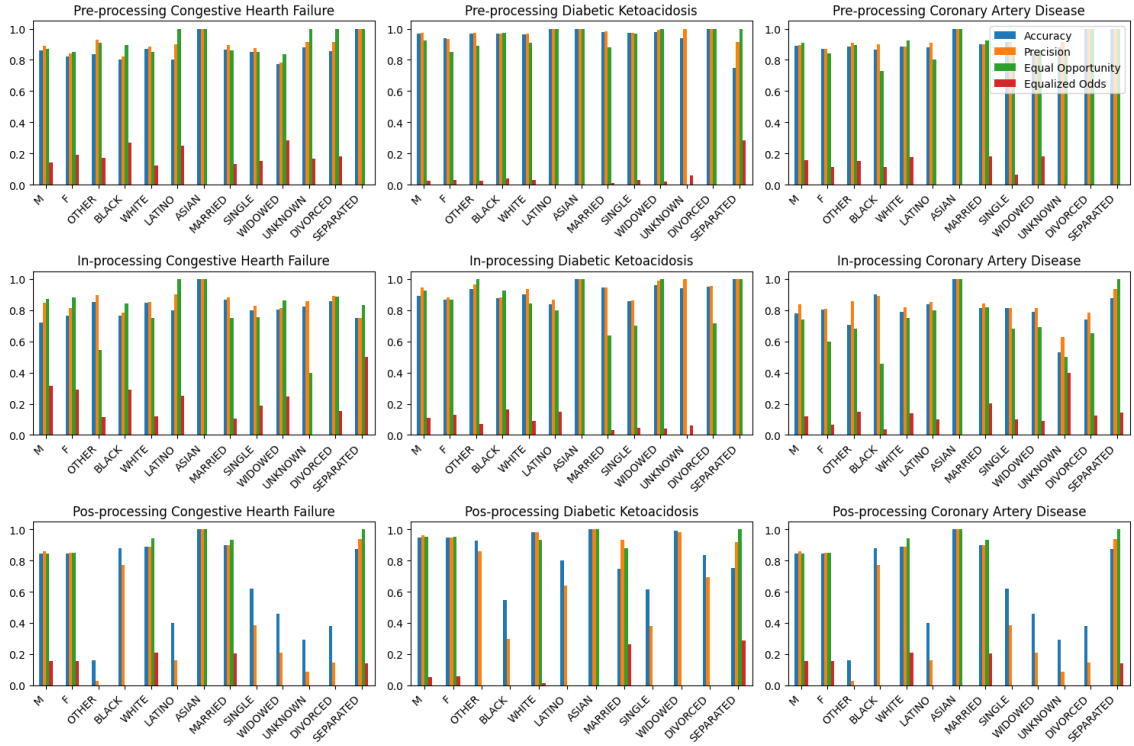


**Figure 3. Fairness strategies with MLP as base model: pre-processing alters input data, in-processing uses adversarial MLP, and post-processing adjusts outputs.**

Post-processing techniques have proven to be more consistent in maintaining predictive accuracy and promoting fairness. As shown in Table 3, the highest loss of accuracy observed was approximately 5% when predicting congestive heart failure using the SVM and coronary artery disease with both the SVM and MLP network. The lowest loss of accuracy, at 1.7%, was observed when predicting congestive heart failure using the MLP algorithm. This demonstrates how the choice of algorithm and data can significantly impact predictive accuracy. For example, switching from SVM to MLP resulted in a 3.2% improvement in predictions.

| | **GBC** | **SVM** | **MLP** |
|---|---|---|---|
| **Diabetic Ketoacidosis** | -1.8% | -3.1% | -2.7% |
| **Congestive Heart Failure** | -3% | -5% | -1.7% |
| **Coronary Artery Disease** | -4.6% | -5% | -5% |

**Table 3. Accuracy variation between original data and post-processing approach.**

Regarding fairness metrics, the MLP network showed the greatest improvement, with an approximately 96.8% increase in the diagnosis rate of diabetic ketoacidosis. In contrast, the smallest improvement—virtually nonexistent—was observed in the prediction of the same condition by the SVM algorithm, as shown in Table 4. To illustrate how these values were calculated, a relevant example is the prediction of heart failure by SVM. In this case, the disparity in the Equal Opportunity metric between sexes decreased from 5% to 1%, representing an 80% reduction after applying the method.

Overall, the results indicate a consistent and noticeable improvement, with fairness metric values converging to a similar range. This improvement occurs with a relatively small loss in predictive performance, which can be a determining factor in contexts where bias mitigation is a higher priority than maximizing model performance. In such scenarios, minimizing the impact of incorrect predictions becomes a crucial aspect.

| | **GBC** | **SVM** | **MLP** |
|---|---|---|---|
| **Diabetic Ketoacidosis** | -83.75% | 0% | -96.8% |
| **Congestive Heart Failure** | -77.5% | -80% | -96.6% |
| **Coronary Artery Disease** | -90.6% | -96.5% | -96.2% |

**Table 4. Variation of Equal Opportunity metric between original data and post-processing approach.**

The pre-processing methods presented some peculiarities, primarily the low impact on accuracy compared to models without fairness considerations. As shown in Table 5, the losses were less than 2% across all the variations tested. This can be attributed to the approach focusing on introducing a new processing step in the data, rather than modifying the predefined models or their training, as seen in the other two cases. This means that when presenting modified data to the same model, the variations pertain to characteristics that might not necessarily affect the predicted classes, as the classes remain unchanged, and only the data used to identify patterns and trends for the predictions differ. The individual characteristics could be maintained to a sufficient extent so that promoting fairness would not compromise the uniqueness of these occurrences.

| | GBC | SVM | MLP |
|---|---|---|---|
| **Diabetic Ketoacidosis** | -1.8% | -1% | -1.7% |
| **Congestive Heart Failure** | -1.1% | -1% | -1.7% |
| **Coronary Artery Disease** | -1% | -1% | -1% |

**Table 5. Accuracy variation between original data and preprocessing approach.**

However, the fairness metrics did not indicate as good performance as the other methods presented earlier. There were positive interactions, as seen in the case of predicting congestive heart failure, where the two metrics showed improvement in general cases. The best case was 86% for Equal Opportunity in the MLP network, while the worst case was just 28% when the Equalized Odds metric was examined with the Gradient Boosting algorithm (refer to Table 6). Moreover, there was a negative impact on the metrics when processing application, but in a more pronounced manner. For instance, when training an MLP network using the dataset for diagnosing diabetic ketoacidosis, there was a worsening of 0.7% in the original data set, which increased to 4% after applying correlation removal. This variation was not observed in any of the other tests.

| | GBC | SVM | MLP |
|---|---|---|---|
| **Diabetic Ketoacidosis** | -45.8% | -44.4% | -85.4% |
| **Congestive Heart Failure** | -27% | -71.4% | -59.6% |
| **Coronary Artery Disease** | 0% | -84.1% | -57.1% |

**Table 6. Variation of Equalized Odds metric between original data and preprocessing approach.**

## 6. Conclusions

This research explores the trade-offs between fairness and performance in machine learning within the healthcare sector, where algorithmic decisions can significantly impact diagnoses, prognoses, and the allocation of medical resources. Three approaches to fairness were evaluated to analyze their effects on predictive performance and equity in outcomes.

Ensuring fairness in healthcare models is crucial to avoid biased predictions that can create disparities in patient care. For instance, ICU admission models trained on unbalanced data may underestimate risk for certain demographic groups, while transplant prioritization algorithms can perpetuate existing inequalities. Among the fairness interventions assessed, post-processing approach achieved the greatest bias reduction but at the cost of predictive performance, while pre-processing better preserved accuracy with only minor deviations. When comparing classifiers, neural networks were the most affected by fairness techniques, particularly in in-processing, which caused the largest performance drop. However, bias reduction was consistent across different diagnoses, indicating that fairness interventions did not disproportionately impact specific disease classifications.

This study provides practical insights into the trade-offs between fairness and predictive performance in machine learning models for healthcare, highlighting the strengths and limitations of different bias mitigation approaches. By demonstrating how post-processing techniques achieve the most significant bias reduction at the cost of performance, while pre-processing methods better preserve predictive accuracy, these findings

offer valuable guidance for developing fairer medical AI systems. Additionally, the observed uniformity in bias reduction across different diagnoses suggests that fairness interventions can be broadly effective without introducing new disparities. Future research should explore these methods in diverse datasets, refine mitigation strategies through parameter tuning, and assess the applicability of multi-class prediction models instead of the binary approach used here. Expanding fairness-aware modeling in this direction could enhance interpretability, simplify testing, and ultimately contribute to more transparent and equitable AI-driven healthcare decisions.

## 7. Acknowledgements

We thank you the financial support of Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP (Grant 2021/14725-3) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

## References

Ay, S., Cardei, M., Meyer, A.-M., Zhang, W., and Topaloglu, U. (2024). Improving equity in deep learning medical applications with the gerchberg-saxton algorithm. *Journal of Healthcare Informatics Research*, 8(2):225–243.

Barocas, S., Hardt, M., and Narayanan, A. (2018). Fairness and machine learning. fairmlbook. org, 2019.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., et al. (2019). Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., and Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*.

Chen, J., Berlot-Attwell, I., Hossain, S., Wang, X., and Rudzicz, F. (2020). Exploring text specific and blackbox fairness algorithms in multimodal clinical nlp. *arXiv preprint arXiv:2011.09625*.

Correa, R., Shaan, M., Trivedi, H., Patel, B., Celi, L. A. G., Gichoya, J. W., and Banerjee, I. (2022). A systematic review of 'fair'ai model development for image classification and prediction. *Journal of Medical and Biological Engineering*, 42(6):816–827.

Ferrara, C., Sellitto, G., Ferrucci, F., Palomba, F., and De Lucia, A. (2024). Fairness-aware machine learning engineering: how far are we? *Empirical software engineering*, 29(1):9.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Jui, T. D. and Rivas, P. (2024). Fairness issues, current approaches, and challenges in machine learning models. *International Journal of Machine Learning and Cybernetics*, pages 1–31.

Kakadiaris, A. (2023). Evaluating the fairness of the mimic-iv dataset and a baseline algorithm: Application to the icu length of stay prediction. *arXiv preprint arXiv:2401.00902*.

Kurbatskaya, A., Jaramillo-Jimenez, A., Ochoa-Gomez, J. F., Brønnick, K., and Fernandez-Quilez, A. (2023). Assessing gender fairness in eeg-based machine learning detection of parkinson's disease: A multi-center study. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pages 1020–1024. IEEE.

Li, J. and Li, G. (2025). Triangular trade-off between robustness, accuracy, and fairness in deep neural networks: A survey. *ACM Comput. Surv.*, 57(6).

Luo, Y., Tian, Y., Shi, M., Pasquale, L. R., Shen, L. Q., Zebardast, N., Elze, T., and Wang, M. (2024). Harvard glaucoma fairness: a retinal nerve disease dataset for fairness learning and fair identity normalization. *IEEE Transactions on Medical Imaging*.

Malone, B., Garcia-Duran, A., and Niepert, M. (2018). Learning representations of missing data for predicting patient outcomes. *arXiv preprint arXiv:1811.04752*.

Meng, C., Trinh, L., Xu, N., Enouen, J., and Liu, Y. (2022). Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *Scientific Reports*, 12(1):7166.

Oliveira, G. M. M. d., Brant, L. C. C., Polanczyk, C. A., Biolo, A., Nascimento, B. R., Malta, D. C., Souza, M. d. F. M. d., Soares, G. P., Xavier Junior, G. F., Machline-Carrion, M. J., et al. (2020). Cardiovascular statistics–brazil 2020. *Arquivos Brasileiros de Cardiologia*, 115:308–439.

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

Raza, S. (2023). Connecting fairness in machine learning with public health equity. In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, pages 704–708. IEEE.

Sattigeri, P., Hoffman, S. C., Chenthamarakshan, V., and Varshney, K. R. (2019). Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5):3–1.

Taber, P., Armin, J. S., Orozco, G., Del Fiol, G., Erdrich, J., Kawamoto, K., and Israni, S. T. (2023). Artificial intelligence and cancer control: toward prioritizing justice, equity, diversity, and inclusion (jedi) in emerging decision support technologies. *Current Oncology Reports*, 25(5):387–424.

Xu, D., Yuan, S., Zhang, L., and Wu, X. (2019). Fairgan+: Achieving fair data generation and classification through generative adversarial nets. In *2019 IEEE international conference on big data (Big Data)*, pages 1401–1406. IEEE.

Yeom, S. and Tschantz, M. C. (2021). Avoiding disparity amplification under different worldviews. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 273–283.

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.