Predicting Mortality and Algorithmic Fairness of ICU Patients using MIMIC-IV

UTRECHT UNIVERSITY

APPLIED DATA SCIENCE

Name: *Afxentis Nikolaos*

Supervisors: *Giachanou Anastasia, Anadria Daniel*

Student Number: *1167138*

ABSTRACT

Predicting mortality for ICU patients while ensuring fairness across different demographic groups is a multifactorial issue. This study aims to address this challenge by leveraging the Medical Information Mart for Intensive Care (MIMIC-IV) dataset to develop robust machine learning models. The study compares neural network and logistic regression models using both a comprehensive set of predictors and a subset of the most significant predictors. Bias mitigation techniques, including reweighting and threshold modification, were applied to address disparities in model performance. Results indicate that while overall accuracy was high, significant biases were observed, particularly against Asian patients and Medicaid insurance holders. The logistic regression model trained on a balanced dataset and adjusted through threshold modification emerged as the optimal choice, achieving minimal inequalities across subgroups while maintaining high accuracy and F1 scores for mortality prediction. These findings underscore the need for continuous evaluation and advanced bias mitigation strategies to ensure equitable healthcare outcomes.

Key Words: Mortality, clinical data, imbalanced dataset, fairness, bias, sensitive attributes, vital signs, laboratory measurements, significant predictors, feature selection, neural network model, logistic regression model, SMOTE, equalized odds, accuracy, F1 score, accuracy, reweighting, threshold modification, ROC curve, calibration plot.

TABLE OF CONTENTS

INTRODUCTION

The Intensive Care Unit (ICU) specializes in treating severely ill patients with essential life-saving interventions such as continuous monitoring of the patient's condition, advanced technologies, and trained staff. This comprehensive approach has been shown to significantly improve patient outcomes by preventing the progression of severe illnesses (Kane et al., 2007). Despite these advancements, predicting patient mortality in the ICU remains a critical area of research. Accurate mortality predictions are vital for optimizing patient care and delivering the most appropriate treatment plans. Ensuring these predictions are fair across diverse patient populations is important for establishing unbiased healthcare and reducing disparities in treatment outcomes.

The rapid advancement of technology and the proliferation of patient health data have necessitated the development of sophisticated machine learning models to accurately predict patient mortality based on specific clinical indicators. These models aim to provide timely and reliable predictions that can assist clinicians in making informed decisions, improving patient care and resource allocation (A. E. W. Johnson et al., 2023). Researchers have primarily focused on enhancing metrics such as accuracy and specificity in these predictive models. However, there has been a notable oversight regarding the potential biases these models may exhibit towards certain demographic groups (Obermeyer et al., 2019). This lack of consideration for fairness and equity in model development can lead to disparities in healthcare outcomes, particularly for protected attributes such as race, gender, and type of insurance.

The aim of this study is to address the dual challenges of predicting mortality in ICU patients who stay for more than 24 hours and mitigating potential algorithmic biases towards protected groups. For the analysis, the Medical Information Mart for Intensive Care (MIMIC-IV) dataset has been used, which includes both structured and unstructured patient data. This research aims to develop robust predictive models that not only achieve high accuracy but also ensure fairness across diverse patient populations.

LITERATURE REVIEW

Forecasting patient outcomes in the ICU is a crucial yet challenging aspect of critical care research. Using multifactorial scoring systems to predict patient outcomes in the ICU has become essential in critical care by offering vital insights into disease severity and guiding long-term treatment strategies. These scoring systems, such as the Acute Physiology and Chronic Health Evaluation (APACHE III) (Knaus et al., 1991), the Sequential Organ Failure Assessment (SOFA) (Antonelli et al., 1999) and the Simplified Acute Physiology Score SAPS II (Gall, 1993), are commonly employed to assess disease severity and mortality risk. These scoring systems were developed through the collaborative efforts of experts in clinical medicine and statistical analysis. Each system was designed with a specific objective and is meant to complement the others. For example, the APACHE III and SAPS are specifically tailored to assess the severity of a patient's condition and predict mortality within the first 24 hours of ICU admission (Pattalung et al., 2021). In general, scoring systems provide valuable insights into the condition of patients in the ICU while they can be used as indicators for variable selection during the analysis.

Such an analysis is imminent due to the abundance of clinical data and the continuous development of technology. Several machine learning algorithms emerged during the last few years aiming to provide insights and to perform accurate predictions. Pang et al. (2022) established and compared multiple machine learning models using the MIMIC-IV database to predict ICU mortality risk. By integrating the Acute Physiology Score III (APS III) (Haq et al., 2014) and the Logistic Organ Dysfunction Score (LODS) (Heldwein et al., 2011) as input features, the study assessed four machine learning models: (a) XGBoost, (b) Logistic Regression, (c) Support Vector Machine (SVM) and (d) Decision Tree. The XGBoost model outperformed others, achieving an AUC of 0.918 which indicates a highly predictive ability. This research highlights the importance of utilizing physiological scores to improve prediction performance and assist clinicians in assessing mortality risk, especially for patients with uncertain survival outcomes. However, the study does not address potential algorithmic biases that may affect certain groups based on protected attributes such as gender, age, ethnicity, and type of insurance. These biases can lead to discrepancies in the accuracy and fairness of predictions resulting in inequitable healthcare outcomes.

In data science, the terms bias and unfairness are closely related and often used interchangeably. However, understanding different types of bias and addressing them effectively is crucial. The study by Chen et al. (2024) provides an extensive review of bias in AI models developed using electronic health records (EHRs). This systematic review defines six major types of bias: algorithmic, confounding, implicit, measurement, selection, and temporal. It identifies various strategies for detecting and mitigating these biases, emphasizing the importance of fairness in AI model development. Key contributions include the evaluation of bias detection methods and bias mitigation techniques, categorized into three stages: (a) preprocessing, (b) in-processing, and (c) post-processing. This paper specifically focuses on algorithmic bias and explores mitigation techniques utilizing these strategies to enhance the fairness and accuracy of machine learning models.

The development of advanced machine learning models and the implementation of bias mitigation algorithms are crucial to ensure that predictions are both fair and accurate across various demographic groups. Additionally, questions have arisen regarding the relationship between the accuracy of these models and the impact of implementing bias mitigation algorithms. Pattalung et al. (2021) and Meng et al. (2022) both focus on predicting mortality in ICU patients using deep machine learning techniques, although they address different considerations of model performance and fairness. Pattalung et al. (2021) leveraged Recurrent Neural Networks (RNNs) to achieve high predictive accuracy (AUCs between 0.87 and 0.91) and utilized SHapley Additive exPlanations (SHAP) for interpretability, ensuring that the factors influencing mortality predictions were transparent and clinically relevant. In contrast, Meng et al. (2022) emphasized the importance of fairness alongside accuracy. Their study identified the IMV-LSTM model as the best performer with an AUC of 0.955, while also highlighting biases and treatment disparities across demographic groups. By evaluating fairness using AUC metrics across these groups, they demonstrated a minimum AUC of 0.918 and a macro-average AUC of 0.954, showing fair predictive performance. Both studies underscore the importance of interpretability and fairness in machine learning models, but Meng et al. (2022) particularly highlights the critical issue of demographic biases, stressing the need for models that are both accurate and unbiased.

Objectives

Many decisions influence the fairness of model outcomes, including the selection of features, model architecture, and bias mitigation techniques. Previous research has not extensively explored the impact of feature selection on the trade-off between performance and fairness. Investigating how different predictors and mitigation strategies affect the accuracy and equity of machine learning models is vital. This study seeks to address these gaps by examining the relationship between predictor selection and model performance, the effectiveness of bias mitigation techniques, and the susceptibility of different patient groups to algorithmic bias.

Research Question:

How effectively machine learning models predict mortality in ICU patients, while mitigating potential biases according to some patients' attributes?

Sub-questions:

a) How does the selection of predictors affect the potential biases across the patients' attributes?

b) How effective are bias mitigation techniques across different demographic groups?

c) Are there groups of patients that are more prone to algorithmic bias?

DATA

2.1 Data Description

In this study, we use MIMIC-IV, which is a publicly accessible database from the Beth Israel Deaconess Medical Center, covering patient data from 2008 to 2019. MIMIC-IV includes comprehensive information on over 60,000 ICU admissions, comprising more than 40,000 unique patients. It includes comprehensive information such as patient demographics, vital signs, laboratory measurements, medication details, clinical notes, procedures, treatments, diagnoses, and patient outcomes. This dataset is designed to facilitate research by providing accessible, structured and unstructured data for developing predictive models and critical care studies (A. E. W. Johnson et al., 2023).

2.2 Pre-processing

Prior to initiating the analysis, several data cleaning and processing steps were conducted. This study specifically targets ICU patients with hospital stays exceeding 24 hours. The time component is essential, as it allows for the collection of more stable and significant clinical data, accurately reflecting ongoing critical care and interventions. Although some patients had multiple ICU admissions, we focused on selecting only their last admission. This approach may introduce some selection bias into analysis, but this is not in the scope of this study. Focusing on a single ICU admission per patient, despite some having multiple admissions, aims to maintain dataset consistency and avoid potential confounding factors. This technique simplifies the analysis and ensures that each patient is represented equally, preventing any one patient from disproportionately influencing the results. Afterwards, variables were categorized into three distinct groups: demographic attributes, vital signs, and laboratory measurements. In the end, after implementing these steps, the dataset contains 12048 rows (number of patients) and 5102 columns (number of variables).

2.3 Feature Engineering

First, the race variable, which is one of the sensitive attributes of this study, was categorized into five distinct groups: (a) Asian (n = 369), (b) Black/African American (n = 1108), (c) Hispanic/Latino (n = 431), (d) White (n = 8068), and (e) Other (n = 2072). This grouping was necessary because the original variable contained subgroups with very few individuals, which could introduce noise and reduce the accuracy of the analysis. The second transformation addressed the missing values in some patient measurements. In this study, no imputation methods were employed to handle the missing clinical data in vital signs and laboratory measurements; instead, NA values were replaced with 0. This approach was chosen to ensure a complete dataset for analysis. However, it is important to note that this method can introduce imputation bias, particularly if the missingness is not under the assumption of missing completely at random.

As previously noted, the abundance of variables poses several challenges to the analysis. The three most principal issues are: (a) Multicollinearity: High correlation among predictor variables can distort the statistical significance of individual predictors decreasing their individual effects. (b) Model Complexity: An increase in the number of variables increases the complexity of the model, thereby reducing its interpretability and making it more difficult to derive clear results. (c) Computational Cost: High dimensionality in data significantly increases the computational time required for model processing, which is detrimental when quick decision-making is essential. Taking all the above into consideration, the analysis is conducted initially using 31 variables closely related to predicting mortality according to Pattalung et al. (2021) and Meng et al. (2022). In Table 1, the names and types of predictors are presented.

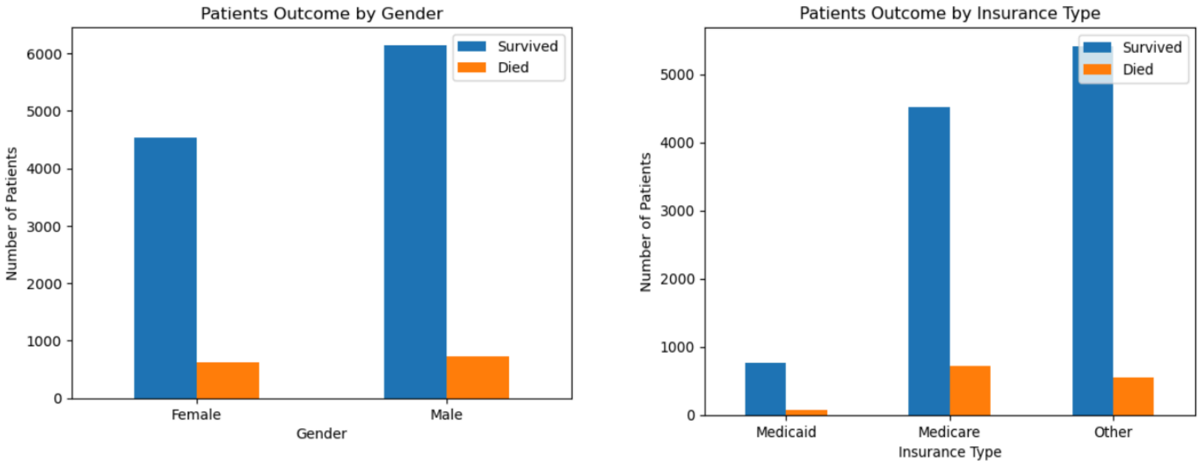
Table 1: Names and types of all 31 predictors

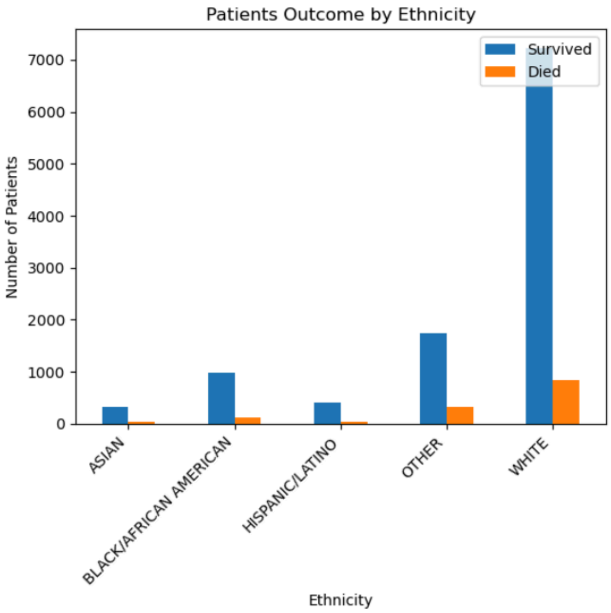
Category	Names of Predictors
Demographic attributes	Gender, Race, Age, Insurance, Admission location, First careunit, Last careunit, Los
Vital Signs	Arterial Blood Pressure diastolic, Arterial Blood Pressure mean, Arterial Blood Pressure systolic, Heart Rate, Oxygen Saturation, Temperature, ART BP Diastolic, ART BP Mean, ART BP Systolic, Bladder Pressure, Blood Temperature CCO (C), EtCO2, PAR-Activity, PAR-Circulation, PAR-Oxygen saturation, PAR-Remain sedated, PAR-Respiration
Laboratory Measurements	Calcium Total, Cholesterol Ratio (Total/HDL), 24hr Creatinine, 24hr Protein, 25-OH Vitamin D

2.4 Exploratory Data Analysis (EDA)

The exploratory data analysis (EDA) phase is a critical component in uncovering the hidden patterns and relationships within the dataset. This phase entails a thorough and systematic examination of the dataset to identify key characteristics and ensure data quality before advancing to more complex modeling and inferential analysis. In this study, the dataset comprises a range of clinical and demographic variables pertinent to ICU patients. The analysis focuses on protected variables such as age, gender, and insurance status.

Figure 1: Bar plots of mortality based on gender, type of insurance and race respectively





By examining Figure 1, it becomes evident that there exists an imbalance between individuals who survived and those who did not across all subgroups of the protected variables. This observation suggests that the models employed in this study are more likely to overfit over the majority class. A significant concern arising from this imbalance is the potential for bias towards certain subgroup categories.

Table 2: Numbers and percentages of patient mortality across different sensitive attributes

Mortality			
Sensitive Attributes		No (%)	Yes (%)
Gender	Male	6147 (89.4%)	729 (10.6%)
	Female	4544 (87.8%)	628 (12.2%)
Race	Asian	328 (88.9%)	41 (11.1%)
	Black/African American	990 (89.3%)	118 (10.7%)
	Hispanic/Latino	401 (93%)	30 (7%)
	Other	1748 (84.4%)	324 (15.6%)
	White	7224 (89.5%)	844 (10.5%)
Insurance	Medicaid	726 (90.3%)	78 (9.7%)
	Medicare	4516 (86.2%)	725 (13.8%)
	Other	5413 (90.7%)	554 (9.3%)

Table 2 shows that male patients have a slightly higher survival rate (89.4%) and a lower mortality rate (10.6%) compared to female patients. Moreover, Hispanic/Latino patients show the best outcomes with the highest survival rate and lowest mortality rate outcomes compared to patients categorized as Other. In terms of the type of insurance, patients with Medicaid or Other types of insurance fare better than those with Medicare, indicating potential disparities based on insurance type. In general, the mortality and survival rates across the sensitive attributes do not differ substantially, but that is not the case based on the number of patients in each group.

However, despite the overall imbalance, patients with Medicaid Insurance are fewer compared to those with Medicare or other types of insurance, increasing the likelihood that the algorithm may develop biases towards the latter groups. Similarly, the distribution of patients by race shows that White patients outnumber the other subgroups suggesting a potential bias towards the White subgroup.

Given these observations, it is crucial to investigate whether the group differences encoded in the dataset could impact the fairness of the models' outcomes trained on this data. Understanding and addressing these biases is vital for ensuring the fairness and accuracy of the predictive models. Analyzing these disparities further can help in developing strategies to mitigate bias and improve the robustness of the models across different subgroups.

METHODOLOGY

This section outlines the steps and algorithms utilized in this analysis. From the initial set of predictors, the objective is to identify the most critical predictors for mortality. Subsequently, the analysis explores how the performance of the predictive models and potential biases change when using all predictor variables compared to using only the most important ones.

3.1 Predictive Models

Generalized Linear Model – Logistic Regression

A Generalized Linear Model (GLM) is a family of models used as an extension of the linear regression model. They allow the dependent variable to have a non-normal distribution. In this research, a GLM is employed to implement binary classification for predicting mortality. The binary nature of the outcome variable (mortality: no/yes) necessitates a model that can handle such data distributions effectively. The specific GLM used in this study is the Logistic Regression model, which is a type of GLM designed for binary outcomes using binary distribution.

Neural Network Model

Neural networks are machine learning algorithms which can recognize complex patterns and relationships in data. Their main components are layers and neurons. Each layer contains neurons which connect to the neurons of the next layer. These connections are called weights. Different layers perform different tasks allowing the model to learn different patterns. Through a process called training, the neural networks adjust the weights of these connections to minimize errors in their predictions.

In this study, four layers have been used. An Input layer containing neurons equal to the number of predictors, two Hidden Layers with 64 and 32 neurons each, and an Output Layer with a single neuron computing the probability of mortality.

3.2 Balancing the Dataset

Synthetic Minority Oversampling Technique (SMOTE)

As it mentioned, dataset imbalance can affect outcome predictions and introduce biases, especially against protected groups. SMOTE (Synthetic Minority Over-sampling Technique) is a method used to address this issue by generating synthetic samples for the minority class. SMOTE works by first identifying the data points that belong to the minority class in the dataset. For each of these points, SMOTE selects a specified number of its nearest neighbors within the minority class (the default is 5). Then, it generates new synthetic samples by choosing a random neighbor and creating a new point somewhere along the line segment that connects the original data point and its neighbor. The exact position of this new synthetic point is determined by a random number between 0 and 1, which ensures that the generated samples are varied (Chawla et al., 2002). By adding these synthetic data points to the dataset, SMOTE increases the representation of the minority class, balancing the class distribution. This approach effectively improves the predictions of the outcome variable for the minority class.

3.3 Bias Mitigation Techniques

Reweighting

Reweighting is an in-processing bias mitigation technique that involves adjusting the weights of the training samples to ensure fair representation of different demographic groups. By assigning higher weights to underrepresented groups and lower weights to overrepresented groups, this technique aims to balance the influence of each group in the training process. This ensures that the model does not disproportionately favor the majority group, thus promoting a more equitable learning process.

Threshold Modification

Threshold modification is a post-processing bias mitigation technique that changes the decision threshold of a classifier to achieve fairer outcomes across different groups. The decision threshold is the value above which a model predicts the positive class. By adjusting this threshold for different demographic groups, the technique aims to balance performance metrics such as the false positive rate and the false negative rate across these groups. This method allows for direct control over the decision criteria to ensure that the classifier treats different groups more equitably, thereby mitigating bias in the model's predictions.

3.4 Evaluation and Fairness Metrics

Accuracy

Accuracy is a commonly used metric for evaluating the performance of a classification model. It measures the proportion of correctly predicted instances out of the total instances in the dataset.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Receiver Operating Characteristic Curve (ROC Curve)

The ROC curve is a graphical representation used to evaluate the performance of a binary classification model. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR).

The Area Under the Curve (AUC) provides a measure model's distinguishability between the two classes, with a higher AUC indicating better performance.

F1 Score

The F1 score is a valuable metric for evaluating the performance of binary classification models. It is especially useful when the dataset is imbalanced, meaning the number of positive instances is much smaller or larger than the number of negative instances.

$$F1\ score\ =\ 2\cdot\frac{\Pr Pr\ e\ cision\cdot Recall}{\Pr Pr\ e\ cision+Recall}$$

Equalized Odds Difference

Equalized odds difference is a fairness metric used to evaluate whether a binary classification model makes equally accurate predictions across different groups. It measures the difference in true positive rates and false positive rates between groups, ensuring that the model's predictions are equally fair and unbiased across all groups. An equalized odds difference close to 0 indicates a fairer model (Hardt et al., 2016).

3.5 Experimental Process

Model Training

To predict mortality using clinical data, two machine learning models were initially applied: a Neural Network (NN) and a Logistic Regression (LR) model. For training the models, two distinct sets of predictors were utilized. The first set contains all 31 predictors available in the dataset, while the second includes only the most significant predictors. Before training the models, categorical variables such as gender, race, insurance, and admission location were converted into numeric format using dummy variables. Afterwards, the dataset was divided into training and test sets, with 80% of the data allocated for training and the remaining 20% for testing. In order to have the same results each time, we set a seed equal to 42.

Both NN and LR models were trained using these predictor sets, and their performance was evaluated on the test set. To assess potential bias towards protected variables, the fairness metric known as equalized odds difference was computed. This initial evaluation provided a baseline understanding of the relationship between models' accuracy and fairness.

Dataset Balancing Using SMOTE

To address class imbalance within the dataset, SMOTE technique was applied. This method generates synthetic samples for the minority class to balance the dataset. Following the application of SMOTE, the NN and LR models were retrained using the balanced dataset. The same procedures as described in the baseline model training were followed, including the conversion of categorical variables to numeric format and the division of the dataset into training and test sets. The objective of retraining the models on the balanced dataset was to evaluate the impact of SMOTE on model accuracy and to investigate whether this technique introduced any bias towards the protected variables.

Mitigating Bias

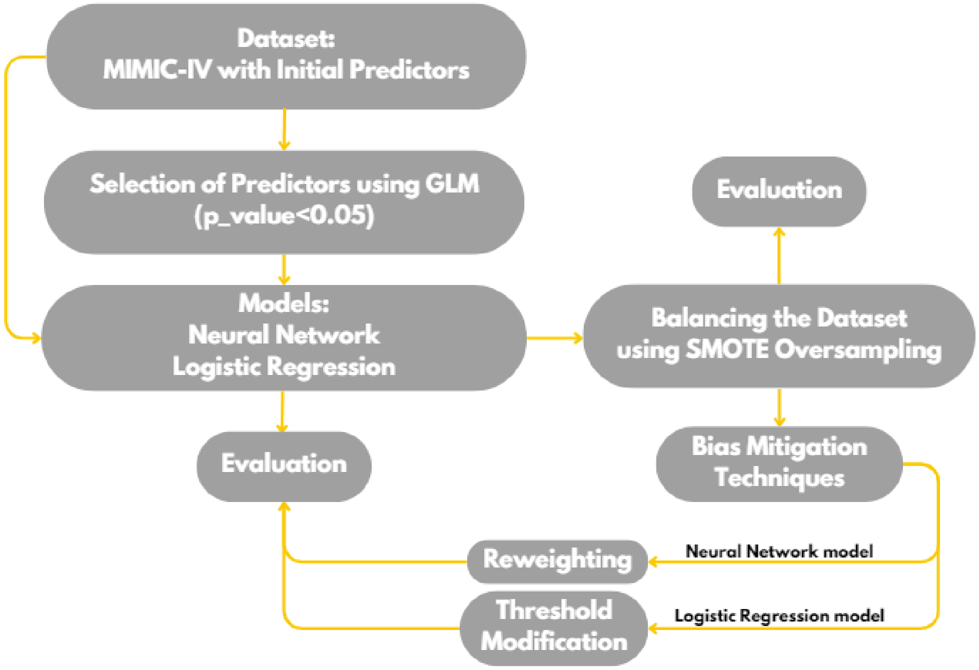
To further address bias, specific bias mitigation techniques were applied to the models. For the neural network models, an in-processing technique reweighting was used. This technique adjusts the weights of the training samples during model training to ensure that the model treats samples from different groups more equitably. For the logistic regression models, a post-processing bias mitigation technique called thresholding was implemented. This method involves adjusting the decision thresholds for different demographic groups after model training to ensure fair outcomes across these groups. The aim is to develop robust predictive models that not only achieve high accuracy but also ensure fairness and reduce bias. Table 3 contains the names and descriptions of the models described in this part.

Table 3: Model name and description

Names	Description
nn_all	Neural network model trained on imbalanced dataset using all predictors
nn_top14	Neural network model trained on imbalanced dataset using only the most significant predictors
lr_all	Logistic regression model trained on imbalanced dataset using all predictors
lr_top14	Logistic regression model trained on imbalanced dataset using only the most significant predictors

nn_smote_all	Neural network model trained on balanced dataset using all predictors
nn_smote_top14	Neural network model trained on balanced dataset using only the most significant predictors
lr_smote_all	Logistic regression model trained on balanced dataset using all predictors
lr_smote_14	Logistic regression model trained on balanced dataset using only the most significant predictors
nn_smote_fair_all	Neural network model trained on balanced dataset + reweighting using all predictors
nn_smote_fair_top14	Neural network model trained on balanced dataset + reweighting using only the most significant predictors
lr_smote_fair_all	Logistic regression model trained on balanced dataset + thresholding using all predictors
lr_smote_fair_top14	Logistic regression model trained on balanced dataset + thresholding using only the most significant predictors

Figure 2: Flowchart of the Experimental Process



RESULTS

4.1 Predictor Selection

To identify the most significant predictors for our analysis, we employed the GLM for binary classification. The primary objective of this model is to determine which predictors have the most significant impact on mortality, specifically those with a p-value less than 0.05. The GLM is trained on an initial set of predictors (31 predictors), and its output highlights the significance of each predictor in forecasting patient mortality. Table 4 contains the names of the predictors that are considered significant (14 predictors).

Table 4: Names and types of only the significant predictors

Category	Names of Predictors
Demographic attributes	Age, Los, Last_careunit Coronary Care Unit (CCU), Last_careunit Medical Intensive Care Unit (MICU), Last_careunit Surgical Intensive Care Unit (SICU), Last_careunit Neuro Surgical Intensive Care Unit (Neuro SICU), Last_careunit Medical Surgical Intensive Care Unit (MICU/SICU), Last_careunit Trauma SICU (TSICU)
Vital Signs	Heart Rate, Oxygen Saturation, Temperature, ART BP Systolic, Bladder Pressure, EtCO2
Laboratory Measurements	-

4.2 Training on the Imbalanced Dataset

As described in the previous section, we trained four machine learning models (2 neural networks and 2 logistic regression models) on the imbalanced dataset. Tables 5, 6, and 7 provide insights for these models in performance and fairness among the demographic groups gender, race, and insurance status.

Table 5: Performance and fairness metrics according to Gender on original dataset

Model	Mortality	F1 scores (Male/Female)	Overall Accuracy	Equalized Odds Difference
nn_all	no	0.93/0.92	0.86	0.007
	yes	0.30/0.29		
nn_top14	no	0.92/0.92	0.86	0.057
	yes	0.32/0.26		
lr_all	no	0.93/0.93	0.87	0.057
	yes	0.11/0.20		
lr_top14	no	0.93/0.93	0.87	0.009
	yes	0.07/0.06		

Table 6: Performance and fairness metrics according to Race on original dataset

Model	Mortality	F1 scores (White/Black-African American/Other/Hispanic-Latino/ Asian)	Overall Accuracy	Equalized Odds Difference
nn_all	no	0.93/0.95/0.89/0.93/0.93	0.86	**0.454**
	yes	0.26/0.39/0.34/**0.48**/**0.00**		
nn_top14	no	0.93/0.94/0.89/0.93/**0.55**	0.86	0.223
	yes	0.28/0.36/0.28/0.35/**0.20**		
lr_all	no	0.94/0.94/0.90/0.94/0.94	0.87	0.128
	yes	0.14/0.08/0.21/0.00/**0.00**		
lr_top14	no	0.93/0.94/0.90/0.94/0.95	0.87	0.041
	yes	0.08/0.00/0.07/0.00/**0.00**		

Table 7: Performance and fairness metrics according to Insurance on original dataset

Model	Mortality	F1 scores (Other/Medicare/Medicaid)	Overall Accuracy	Equalized Odds Difference
nn_all	no	0.93/0.92/0.93	0.86	0.132
	yes	0.20/0.37/0.32		
nn_top14	no	0.93/0.91/0.94	0.86	0.153
	yes	0.28/0.32/0.17		
lr_all	no	0.94/0.92/0.93	0.87	0.145
	yes	0.05/0.24/0.00		
lr_top14	no	0.94/0.92/0.94	0.87	**0.050**
	yes	0.04/0.09/0.00		

In terms of overall accuracy, all the models achieve high accuracy around 0.87, indicating strong overall performance. However, despite the high accuracy, there are disparities in the F1 scores, particularly for minority classes. While the F1 scores for the majority class, patients that survived, are mostly over 0.90, the scores for minority classes are notably lower because of the class imbalance. Furthermore, there is notable variation in F1 scores, especially among demographic subgroups of race attribute. This inserts bias to the models' predictions making the models more or less favorable to patients with specific characteristics.

According to predictor selection, models using all predictors tend to have higher F1 scores compared to those using only the most significant predictors. This suggests that including a broader range of predictors enhances the models' ability to accurately classify different classes. Notably, *lr_top14* model shows the lowest F1 scores for minority classes across all demographic attributes, with values ranging from 0 to 0.1, indicating that this model is less effective in handling minority class predictions. Moreover, neural network models generally outperform logistic regression models, especially when only the most significant predictors are selected and predictions over minority class are considered.

As fairness is concerned based on gender, the equalized odds difference metric is close to 0 in all cases. So, there is no substantial algorithmic bias between male and female patients. Such findings are to be expected given the distributions of the outcome across the protected attributes in the training data. However, when considering the race attribute, the fairness metric for the *nn_all* model is 0.45 which shows moderate bias towards some subgroups. The F1 scores for the Hispanic/Latino and Asian subgroups in this model are 0.48 and 0, respectively, highlighting high variance and potential bias. The Asian subgroup has the lowest F1 scores across all models and regardless of the selection of predictors. Additionally, the equalized odds difference experiences some variations: neural network models, despite their relatively better performance, have higher EOD values compared to logistic regression models. Models with fewer predictors tend to have lower equalized odds difference scores compared to those using all predictors. Finally, the EOD values for insurance status are considerably low, with the lowest observed value being 0.05 in the *lr_top14* model. This indicates minimal bias in model performance related to insurance type.

4.3 Training after SMOTE Oversampling

Using SMOTE Oversampling technique, the dataset is balanced over the outcome variable. Tables 8, 9, and 10 provide insights for performance and fairness of the models among the demographic groups gender, race, and insurance status after balancing the dataset by generating data for minority class.

Table 8: Performance and fairness metrics according to Gender on balanced dataset

Model	Mortality	F1 scores (Male/Female)	Overall Accuracy	Equalized Odds Difference
nn_smote_all	no	0.87/0.85	0.77	0.090
	yes	0.38/0.31		
nn_smote_top14	no	0.84/0.82	0.73	0.025
	yes	0.39/0.37		
lr_smote_all	no	0.92/0.92	0.86	0.035
	yes	0.34/0.31		
lr_smote_top14	no	0.90/0.87	0.81	0.041
	yes	0.46/0.42		

Table 9: Performance and fairness metrics according to Race on balanced dataset

Model	Mortality	F1 scores (White/Black-African American/Other/Hispanic-Latino /Asian)	Overall Accuracy	Equalized Odds Difference
nn_smote_all	no	0.86/0.85/0.84/0.89/0.95	0.77	**0.369**
	yes	0.34/0.29/0.43/0.32/**0.22**		
nn_smote_top14	no	0.83/0.85/0.81/0.86/**0.81**	0.73	0.389
	yes	0.37/0.37/0.43/0.47/0.21		
lr_smote_all	no	0.93/0.95/0.88/0.93/0.95	0.86	0.333
	yes	0.31/0.43/0.37/0.25/**0.00**		
lr_smote_top14	no	0,89/0.90/0.87/0.87/0.86	0.81	**0.298**
	yes	0.44/0.44/0.49/0.44/**0.26**		

Table 10: Performance and fairness metrics according to Insurance on balanced dataset

Model	Mortality	F1 scores (Other/Medicare/Medicaid)	Overall Accuracy	Equalized Odds Difference
nn_smote_all	no	0.90/**0.80**/0.94	0.77	**0.469**
	yes	0.34/0.37/**0.17**		
nn_smote_top14	no	0.86/0.79/0.82	0.73	0.083
	yes	0.40/0.37/0.34		
lr_smote_all	no	0.93/0.91/0.93	0.86	**0.367**
	yes	0.23/0.42/**0.00**		
lr_smote_top14	no	0.90/0.87/0.90	0.81	0.158
	yes	0.39/0.49/0.41		

By balancing and training the models on the new dataset, several new findings emerged. The overall accuracy for the NN models decreased by approximately 12%, while for the logistic regression LR models, there was only a slight decrease of 7% for the model that uses only the significant predictors. Models using the most significant predictors experienced a more substantial decrease in overall accuracy.

Regarding F1 scores, an increase was noted in most subgroups with some experiencing a more substantial increase than others. This indicates that bias may have been introduced, improving the models' predictions for only some of the demographic subgroups.

In case of gender attribute, the F1 scores have been increased for both groups without indicating potential bias towards male or female patients. The EOD also remained low, close to 0, meaning there is no indication of gender bias.

According to race, we observed that some subgroups experienced significant increase in F1 scores compared to others for minority class. For example, the F1 score for the Asian patients increased from 0 to 0.22 when using the *nn_smote_all* and from 0 to 0.26 when using the lr_smote_top14 model. An interesting finding is that also a significant increase from 0.55 to 0.81 occurred in the majority class for the Asian group using the *nn_smote_top14* model. Additionally, an increase in EOD has been noted with *lr_smote_top14* to be 87%, while the value of EOD for *nn_smote_all* decreased by 21%. The EOD values range from 0.3 to 0.4, showing potential bias to the models.

For the insurance attribute, we noticed a significant increase in both F1 scores and EOD metrics, especially for models using all predictors. For patients with Medicaid insurance, the F1 score using the *nn_smote_all* model dropped to 0.17, while for other groups, the scores increased. A similar trend was observed for the Medicaid group using the *lr_smote_all* model, where the F1 score remained 0, but the scores for other groups increased considerably. The EOD metric also experienced a substantial increase, particularly for the *nn_smote_all* and *lr_smote_all* models, reaching values of 0.47 and 0.37, respectively. This indicates that models using more predictors are more prone to bias in this case.

4.4 Training after SMOTE Oversampling and Bias Mitigation

In order to address the issue of the inserted bias after balancing the data, two mitigation bias approaches have been conducted. For NN models, it is an in-processing technique called reweighting, while for the LR models a post-processing technique called thresholding. Tables 11, 12 and 13 present the results after implementing these methods.

Table 11: Performance and fairness metrics according to Gender after bias mitigation

Model	Mortality	F1 scores (Male/Female)	Overall Accuracy	Equalized Odds Difference
nn_smote_fair_all	no	0.87/0.88	0.80	0.057
	yes	0.43/0.44		
nn_smote_fair_top14	no	0.82/0.80	0.71	0.009
	yes	0.40/0.38		
lr_smote__fair_all	no	0.88/0.89	0.81	0.133
	yes	0.46/0.43		
lr_smote_fair_top14	no	0.88/0.85	0.79	0.035
	yes	0.47/0.43		

Table 12: Performance and fairness metrics according to Race after bias mitigation

Model	Mortality	F1 scores (White/Black-African American/Other/Hispanic-Latino/Asian)	Overall Accuracy	Equalized Odds Difference
nn_smote_fair_all	no	0.88/0.87/0.85/0.90/0.94	0.80	**0.464**
	yes	0.41/0.42/0.51/0.43/0.33		
nn_smote_fair_top14	no	0.81/0.84/0.79/0.81/0.81	0.71	0.178
	yes	0.38/0.38/0.44/0.37/0.27		
lr_smote_fair_all	no	0.89/0.90/0.84/0.91/0.95	0.81	**0.708**
	yes	0.43/0.48/0.50/0.46/**0.00**		
lr_smote_fair_top14	no	0.87/0.88/0.84/0.85/0.85	0.79	**0.571**
	yes	0.45/0.42/0.50/0.51/0.25		

Table 13: Performance and fairness metrics according to Insurance after bias mitigation

Model	Mortality	F1 scores (Other/Medicare/Medicaid)	Overall Accuracy	Equalized Odds Difference
nn_smote_fair_all	no	0.89/0.84/0.92	0.80	0.314
	yes	0.42/0.46/**0.38**		
nn_smote_fair_top14	no	0.84/0.78/0.81	0.71	0.101
	yes	0.39/0.40/0.35		
lr_smote_fair_all	no	0.88/0.88/0.91	0.81	**0.127**
	yes	0.42/0.48/**0.43**		
lr_smote_fair_top14	no	0.87/0.86/0.83	0.79	0.067
	yes	0.42/0.50/0.39		

After implementing bias mitigation techniques, we observed a further overall increase in F1 scores across nearly all groups of the protected attributes. Patients with Medicaid insurance experienced the most significant improvement. Specifically, using the reweighting technique, the F1 score for the *nn_smote_fair_all* model increased from 0.17 to 0.38. Similarly, using the threshold modification technique, the F1 score for the *lr_smote_fair_top14* model increased from 0 to 0.43. Additionally, the EOD values decreased in each case, with the most significant reduction occurring in the *lr_smote_fair_top14* model, where EOD decreased by 66%.

Regarding the race attribute, the results were more mixed. In the *lr_smote_fair_all* model, the F1 score for the Asian group remained at 0, even after applying bias mitigation techniques. Furthermore, the EOD for this model increased to 0.71, indicating that threshold modification was not effective in addressing bias for the race attribute. Additionally, an increase of 24% in the EOD metric was observed for the *nn_smote_fair_all* model, suggesting that the reweighting method did not adequately mitigate bias among race subgroups.

For the gender category, no substantial difference occurred to the EOD metric with all values remaining close to 0 in each case. Although, in terms of model comparison, it is presented that models using only the most significant predictors have lower EOD values in each protected attribute and despite the type of the model.

4.5 Summary

As infered above, the bias mitigation techniques did not effectively address bias for the race attribute, prompting further exploration of the underlying causes. Calibration plots and ROC-AUC plots can provide valuable insights into this issue.

According to Figures 4 and 6, the AUC scores for Asian patients are lower compared to other groups, indicating poorer discriminatory performance for this subgroup. Additionally, the calibration plots in Figures 5 and 7 reveal substantial deviations from the calibration line for Asian patients compared to other patients, particularly in models utilizing all predictors. This suggests that the predicted probabilities for Asian patients are less reliable, contributing to the observed biases.

When examining the relationship between performance and fairness, it becomes evident that efforts to enhance model performance can also increase the EOD fairness metric. This trade-off suggests that optimizing for overall performance may provoke disparities among protected groups (Figures 8 and 9).

Figure 4: ROC curve of *nn_smote_fair_all* and *nn_smote_fair_top14* respectively

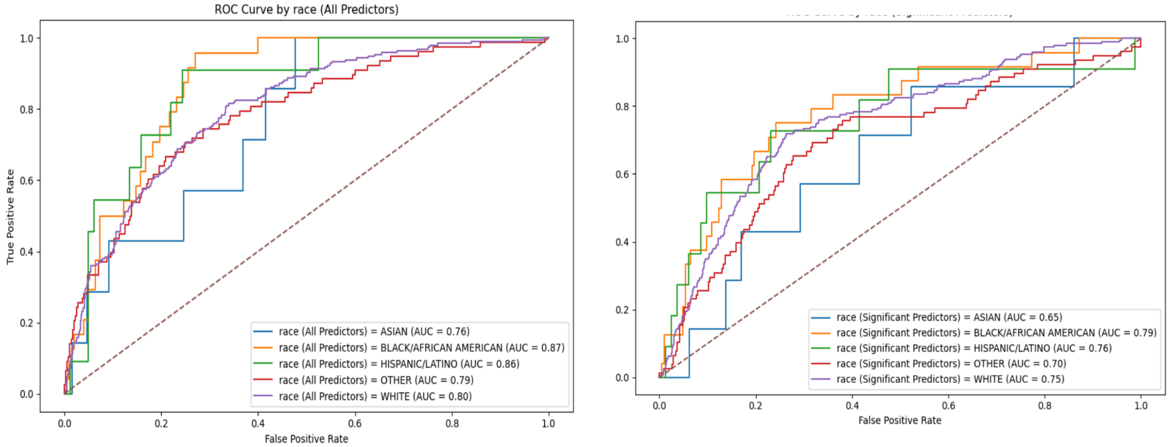


Figure 5: Calibration curve of *nn_smote_fair_all* and *nn_smote_fair_top14* respectively

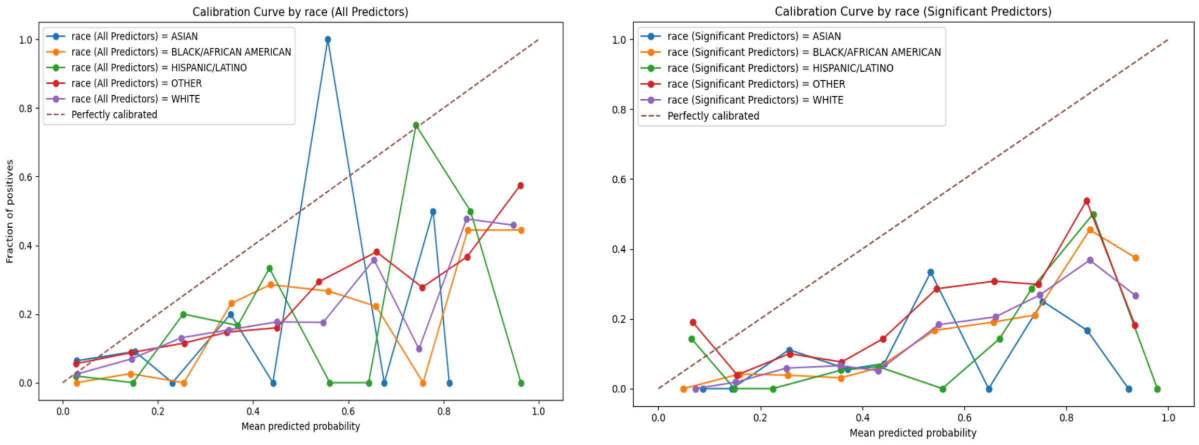


Figure 6: ROC curve of *lr_smote_fair_all* and *lr_smote_fair_top14* respectively

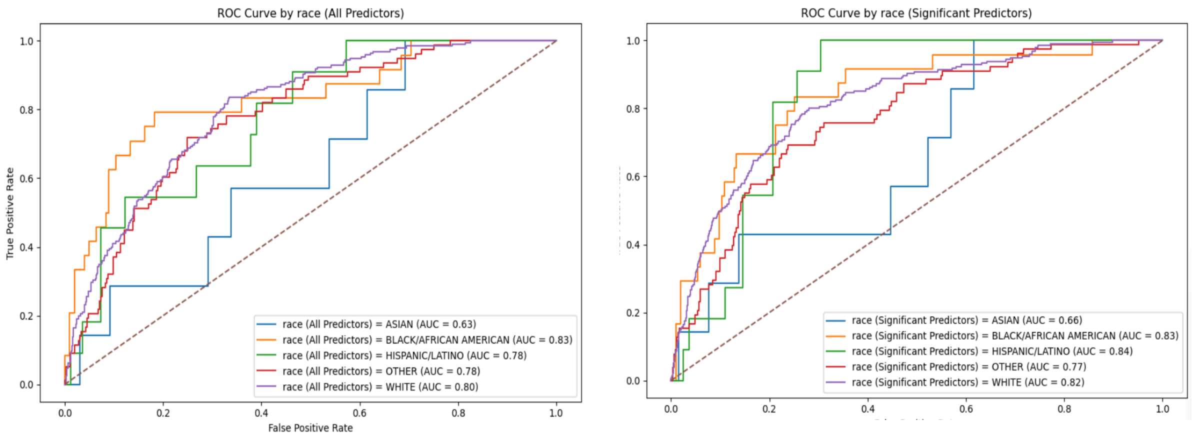


Figure 7: Calibration curve of *lr_smote_fair_all* and *lr_smote_fair_top14* respectively

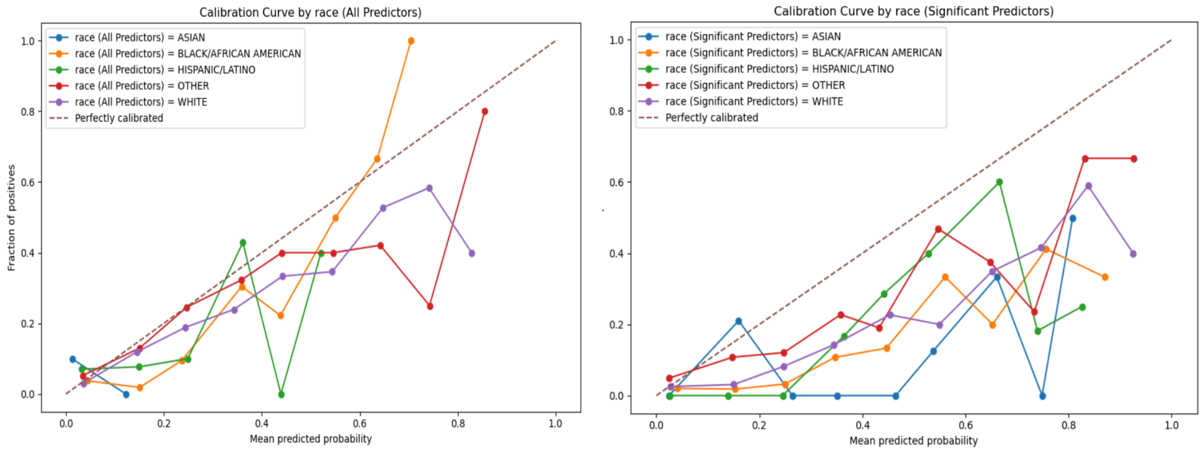


Figure 8: ROC-AUC curve for neural network and logistic regression models

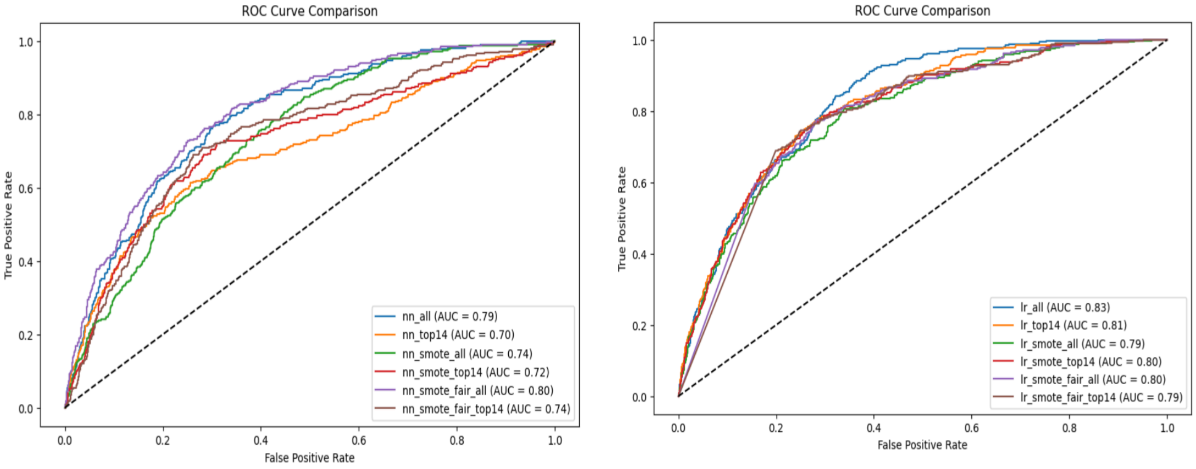
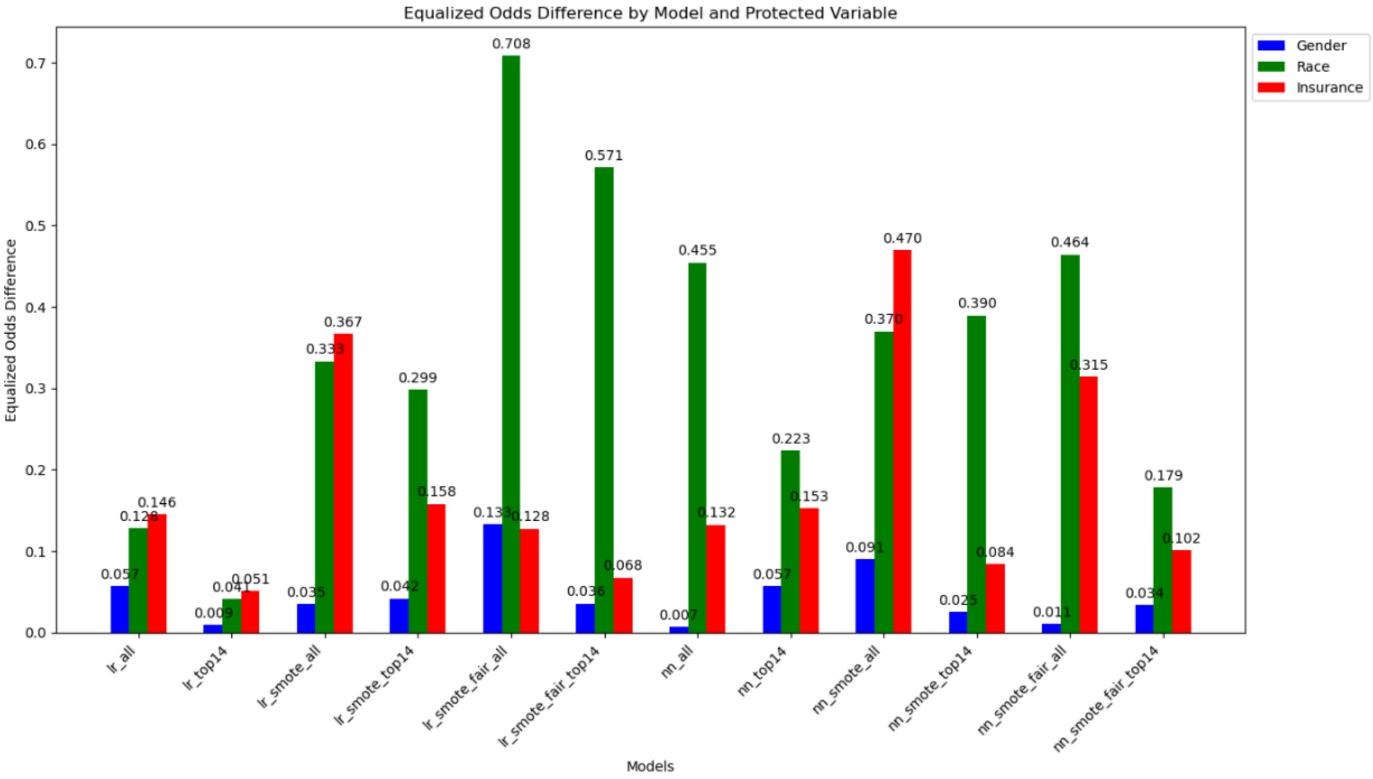


Figure 9: Equalized Odds Difference bar plot for every model



DISCUSSION

The unequal distribution of patients across different protected groups can decrease model performance and introduce algorithmic bias, particularly regarding race and type of insurance. The primary research question investigates how machine learning models can accurately predict mortality in ICU patients while mitigating potential biases. Analysis reveals a trade-off between performance and fairness based on the number of predictors used. Models utilizing all predictors tend to predict patient mortality more accurately, with neural networks outperforming logistic regression models. However, models using only the most significant predictors exhibit less bias towards demographic groups, making bias mitigation techniques more effective.

So, for the question of how the selection of predictors impacts potential biases on patients' sensitive attributes, we can infer that models using a subset of the most significant predictors demonstrated enhanced fairness across demographic groups. These models exhibited reduced bias towards protected attributes such as race and type of insurance, ensuring more equitable treatment and outcomes for different patients. By focusing on the most significant predictors, the models balanced the trade-off between accuracy and fairness more effectively.

Balancing the dataset by generating additional data to improve prediction accuracy for minority classes inadvertently introduced bias, favoring some patient subgroups over others. The gender attribute did not experience significant changes, with EOD values remaining close to 0, indicating no bias between male and female patients. This outcome aligns with the EDA findings, where the distribution of patients by gender was relatively equal. In contrast, race and insurance type attributes showed more pronounced biases. Asian patients were less favored in terms of race, while patients with Medicaid insurance were less favored in terms of insurance type. This is evident from the F1 scores and EOD values for these groups, with the race attribute showing the highest EOD values almost in every case.

Regarding the question of how effective bias mitigation techniques are across different demographic groups, both reweighting and threshold modification successfully mitigated bias for the insurance attribute, particularly for patients with Medicaid insurance. However, these techniques were less successful in mitigating bias for the race attribute, particularly for Asian patients using the threshold modification technique. This means that models fail to treat Asian patients fairly, provoking discrimination towards this group. The F1 scores for Asian patients were substantially lower than the other groups for minority class. So, the models underestimated the risk of mortality for these patients contributing to delaying necessary medical interventions, leading to the progression of diseases or even mortality. To answer the question of whether there are groups of patients more prone to bias, the analysis reveals that certain groups, specifically Asian patients and those with Medicaid insurance, are particularly susceptible to algorithmic bias.This is against the principle that all patients must be treated equally, regardless of their racial background (Beauchamp et al., 2019). Such discrimination not only violates ethical standards but also exposes healthcare institutions to legal risks, including potential lawsuits and regulatory penalties for violating anti-discrimination laws. Ensuring equitable treatment is essential for maintaining trust in the healthcare system and providing high-quality care to all patients.

In this research, to achieve both accurate and unbiased mortality predictions, the neural network model using only the significant predictors, trained on the balanced dataset and adjusted through reweighting emerges as the optimal choice. This model demonstrates minimal inequalities across subgroups while maintaining high accuracy and F1 scores for mortality prediction.

CONCLUSION

The main objective of this study was to predict ICU patient mortality with a focus on fairness across various demographic groups. The findings reveal a trade-off between performance and fairness, with models utilizing all predictors achieving higher accuracy but also exhibiting greater bias. Bias mitigation techniques effectively addressed disparities related to insurance type but were less successful for racial attributes, particularly for Asian patients. This led to underestimation of mortality risk and potential delays in medical interventions for this group. The neural network model, trained on a balanced dataset and adjusted through reweighting, demonstrated the best balance of accuracy and fairness. These results highlight the necessity for more sophisticated bias mitigation approaches that can further address bias. Future research should focus on refining these techniques or finding a different set of predictors increasing the robustness of the models.

Data Availability Statement and Code Availability

Restrictions apply to the availability of these data. Data were obtained from MIMIC-IV and are available at https://physionet.org/content/mimiciv/2.2/ with the permission of PhysioNet. The Jupyter notebooks are available on GitHub on the following link:

(https://github.com/linos98/Predicting-Mortality-and-Algorithmic-Fairness-of-ICU-Patients)

Acknowledgments

I am deeply grateful to the John S. Latsis Public Benefit Foundation for awarding me a scholarship that generously covered my expenses for my studies.

REFERENCES

Kane, R. L., Shamliyan, T. A., Mueller, C., Duval, S., & Wilt, T. J. (2007). The association of registered nurse staffing levels and patient outcomes. Medical Care, 45(12), 1195–1204.

Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L. H., Celi, L. A., & Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data, 10(1).

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

Knaus, W. A., Wagner, D. P., Draper, E. A., Zimmerman, J. E., Bergner, M., Bastos, P. G., Sirio, C. A., Murphy, D. J., Lotring, T., Damiano, A., & Harrell, F. E. (1991). The APACHE III Prognostic System. Chest, 100(6), 1619–1636.

Antonelli, M., Moreno, R., Vincent, J. L., Sprung, C. L., Mendoça, A., Passariello, M., Riccioni, L., & Osborn, J. (1999). Application of SOFA score to trauma patients. Intensive Care Medicine, 25(4), 389–394.

Gall, J. L. (1993). A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA, 270(24), 2957.

Pattalung, T. N., Ingviya, T., & Chaichulee, S. (2021). Feature explanations in recurrent neural networks for predicting risk of mortality in intensive care patients. Journal of Personalized Medicine, 11(9), 934.

Meng, C., Trinh, L., Xu, N., Enouen, J., & Liu, Y. (2022). Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. Scientific Reports, 12(1).

Pang, K., Li, L., Ouyang, W., Liu, X., & Tang, Y. (2022). Establishment of ICU Mortality Risk Prediction Models with Machine Learning Algorithm Using MIMIC-IV Database. Diagnostics, 12(5), 1068.

Haq, A., Patil, S., Parcells, A. L., & Chamberlain, R. S. (2014). The Simplified Acute Physiology Score III is superior to the Simplified Acute Physiology Score II and Acute Physiology and Chronic Health Evaluation II in predicting surgical and ICU mortality in the "Oldest Old." Current Gerontology and Geriatrics Research, 2014, 1–9.

Heldwein, M. B., Badreldin, A. M., Doerr, F., Lehmann, T., Bayer, O., Doenst, T., & Hekmat, K. (2011). Logistic Organ Dysfunction Score (LODS): A reliable postoperative risk management score also in cardiac surgical patients? Journal of Cardiothoracic Surgery, 6(1).

Chen, F., Wang, L., Hong, J., Jiang, J., & Zhou, L. (2024). Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. Journal of the American Medical Informatics Association, 31(5), 1172–1183.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling technique. Journal of Artificial Intelligence Research/the Journal of Artificial Intelligence Research, 16, 321–357.

Hardt, M., Price, E., & Srebro, N. (2016, October 7). Equality of opportunity in supervised learning.

Beauchamp, T. L., & Childress, J. F. (2019). *Principles of biomedical ethics* (8th ed.). Oxford University Press.

APPENDIX

Table 14: Significance of predictors using GLM Model

Name of predictor	P_value	Significant
Age	0.000	yes
Los	0.025	yes
Arterial Blood Pressure diastolic	0.074	no
Arterial Blood Pressure mean	0.785	no
Arterial Blood Pressure systolic	0.396	no
Heart Rate	0.000	yes
Calcium, Total	0.069	no
Cholesterol Ratio (Total/HDL)	0.482	no
Oxygen Saturation	0.000	yes
Temperature	0.000	yes
ART BP Diastolic	0.734	no
ART BP Mean	0.950	no
ART BP Systolic	0.043	yes
Bladder Pressure	0.012	yes
Blood Temperature CCO (C)	0.822	no
EtCO2	0.000	yes
PAR-Activity	0.284	no
PAR-Circulation	0.934	no
PAR-Oxygen saturation	0.897	no
PAR-Remain sedated	0.180	no
PAR-Respiration	0.075	no
24hr Creatinine	1.000	no
24hr Protein	0.214	no
25-OH Vitamin D	0.281	no
Gender_male	0.203	no
Race_BLACK/AFRICAN AMERICAN	0.504	no
Race_HISPANIC/LATINO	0.135	no
Race_OTHER	0.101	no
Race_WHITE	0.678	no
Insurance_Medicare	0.630	no
Insurance_Other	0.929	no
Admission location_CLINIC REFERRAL	0.488	no
Admission location_EMERGENCY ROOM	0.914	no
Admission location_INFORMATION NOT AVAILABLE	0.783	no
Admission location_PACU	0.999	no

Admission location_PHYSICIAN REFERRAL	0.512	no
Admission location_PROCEDURE SITE	0.484	no
Admission location_TRANSFER FROM HOSPITAL	0.957	no
Admission location_TRANSFER FROM NURSING FACILITY	0.883	no
Admission location_WALK IN/SELF REFERRAL	0.677	no
First_careunit_Coronary Care Unit (CCU)	0.493	no
First_careunit_Medical Intensive Care Unit (MICU)	0.512	no
First_careunit_Medical/Surgical Intensive Care Unit (MICU/SICU)	0.818	no
First_careunit_Neuro Intermediate	0.893	no
First_careunit_Neuro Stepdown	0.520	no
First_careunit_Neuro Surgical Intensive Care Unit (Neuro SICU)	0.645	no
First_careunit_Surgical Intensive Care Unit (SICU)	0.092	no
First_careunit_Trauma SICU (TSICU)	0.779	no
Last_careunit_Coronary Care Unit (CCU)	0.000	yes
Last_careunit_Medical Intensive Care Unit (MICU)	0.000	yes
Last_careunit_Medical/Surgical Intensive Care Unit (MICU/SICU)	0.000	yes
Last_careunit_Neuro Intermediate	0.983	no
Last_careunit_Neuro Stepdown	0.792	no
Last_careunit_Neuro Surgical Intensive Care Unit (Neuro SICU)	0.000	yes
Last_careunit_Surgical Intensive Care Unit (SICU)	0.000	yes
Last_careunit_Trauma SICU (TSICU)	0.000	yes