# Classification Analysis of Acute Respiratory Distress Syndrome using preECMO data

Thesis submitted in accordance with the requirements of the University of Glasgow for the degree of **Master of Science** by

**Kavya Kayiparambil Harish**

*1 September 2023*

# ABSTRACT

Acute Respiratory Distress Syndrome (ARDS) is a life-threatening lung injury that has high a mortality rate. A new treatment to improve the disease outcome for ARDS patients is the Extracorporeal membrane oxygenation (ECMO). As ECMO is a complex and high-risk therapy, it should be used for patients who are likely to get benefit from it.

We analysed medical records of 450 ARDS patients who received ECMO treatment. To identify the most important features in the dataset, we used Lasso regression and Boruta algorithm. Both of these methods highlighted lactate, indication 6, and respiratory rate as important factors. We then compared logistic regression, Naive Bayes, decision trees, random forest and neural networks to predict the ECMO candidacy of ARDS patients. The performance of the classifiers were measured using Matthews Correlation Coefficient (MCC), F1 score and accuracy. Random forest achieved highest scores on a complete imbalanced dataset (MCC: 0.27) while neural networks achieved superior performance on the balanced under-sampled dataset (MCC: 0.72).

Our results show that Lasso and Boruta can efficiently identify the important features in this dataset and machine learning algorithms can predict the ECMO candidature with high accuracy.

# List of Tables

# List of Figures

# Contents

# 1 Introduction

Acute Respiratory Distress Syndrome (ARDS) is a critical medical condition characterized by sudden and severe lung inflammation, leading to a significant decrease in oxygen levels in the blood. ARDS can arise from various causes, including pneumonia, sepsis, or severe injuries. The condition is marked by symptoms such as rapid breathing, low blood oxygen levels, and lung infiltrates visible on imaging studies.

A 2016 study found that ARDS occurred in 10.4% of ICU patients and 23% of mechanically ventilated patients across 50 nations, with an overall hospital mortality rate of approximately 40% [3, 22, 10]. Furthermore, the COVID-19 pandemic has led to an upsurge in the number of ARDS cases. There is currently no specific diagnostic test to detect the condition and the inherent heterogeneity of the disease poses significant challenges in understanding, and diagnosing the condition.

Treatment for ARDS focuses on providing mechanical ventilation and supportive care to maintain oxygenation and manage the underlying cause. ARDS can be a challenging condition to manage, and despite advances in critical care, it remains associated with high morbidity and mortality rates. A detailed information about the definition, diagnosis, and epidemiology of ARDS can be found in [17].

Extracorporeal Membrane Oxygenation (ECMO) is an advanced medical therapy that serves as a temporary life support system for individuals with severe cardiac and pulmonary dysfunction [16]. This technique has evolved into a crucial tool for treating patients who are unresponsive to conventional management. ECMO treatment involves the use of a specialized machine that temporarily assumes the function of these vital organs, oxygenating the blood outside the body and removing carbon dioxide, providing the patient's heart and lungs a chance to rest and heal. Although ECMO is a highly complex therapy, its success in saving lives and improving patient outcomes has made it an indispensable tool in critical care settings around the world.

Over the past decade, several clinical trials have been conducted to evaluate the benefits of ECMO compared to conventional management for severe respiratory failure [15]. The results have been mixed, and the efficacy of ECMO can depend on factors such as patient selection criteria, the underlying condition, etc. In some trials,

ECMO has shown to significantly improve survival rates and outcomes compared to conventional therapies. However, in other trials, ECMO has not demonstrated a significant advantage over conventional management, or the benefits may have been limited to certain subgroups of patients. ECMO is a highly specialized and resource-intensive treatment that carries significant risks and complications. Therefore, it should be reserved for patients who are most likely to benefit from the treatment.

In this project we analysed medical records of 450 ARDS patients who have undergone ECMO therapy. Each individual record encompassed a multitude of biomarkers alongside a clinical outcome. The outcome variable has two levels and indicates whether the patient survived the therapy or not.

Healthcare data often contains highly correlated variables since each aspect of the health of an individual is related to their other health aspects. Such multicollinearity in data can cause problems with developing prediction models. Moreover, the lesser the number of predictors the more the possibility of obtaining a simpler and more understandable model. Due to that reason we initiated the analysis with feature selection algorithms. We used hypothesis testing, Lasso regression [20] and Boruta algorithm [13]. Following feature selection, we employed Machine Learning (ML) algorithms for prediction. ML models are widely recognized as a valuable asset in clinical data analysis due to its ability to uncover complex patterns, identify subtle relationships, and extract valuable insights from complex datasets. We implemented five ML algorithms; logistic regression, Naive Bayes, decision trees, random forest and neural network. We chose these methods because they were suitable for our dataset, and we wanted to compare different machine learning techniques for our binary classification problem.

Our results and methods can help medical practitioners in identifying ARDS patients who are likely to benefit from the ECMO therapy. Additionally, our observations provide guidance on which biomarkers medical practitioners should take into account when assessing a patient's suitability for treatment.

## 1.1 Aim

The aim of this project is to investigate the association between specific biomedical markers and the suitability of ARDS patients for ECMO treatment. The investigation

is done using the biomarker data of ARDS patients. The goal is predict whether a patient will survive the ECMO treatment.

## 1.2 Objectives

The main objectives of this thesis are:

- Identifying the suitable statistical method for prediction
- Preparing a model for accurate predictions
- Identifying the features (biomedical markers) that may serve as reliable indicators for predicting ECMO candidacy

## 1.3 Outline of thesis

The thesis is organised as follows: chapter 2 contains a review of techniques used during this project, chapter 3 discusses the methodology used, chapter 4 present the results obtained, and chapter 5 concludes the thesis.

# 2 Background

## 2.1 Variable selection techniques

Variable selection can be understood in two ways: identifying the set of true relevant variables or doing dimension reduction by transforming true variables in to a set of new variables. In this project we focus on the former, since one of the aims of the project is to identify relevant covariates.

### Lasso regression

Lasso regression, also known as L1 regularization, addresses the challenge of multi-collinearity and performs variable selection by adding a penalty term to the regression model based on the sum of the absolute values of the coefficients. Lasso regularization encourages certain coefficients to be exactly zero, effectively eliminating less relevant predictors and leading to a sparse model with only the most influential features retained. This characteristic of Lasso makes it valuable for handling high-dimensional datasets and identifying the most critical variables for accurate predictions and model interpretability.

### Boruta

Boruta is a feature selection algorithm designed to enhance the performance of machine learning models by identifying the most relevant features from a given dataset. It was inspired by the Random Forest algorithm and is particularly useful for datasets with a high number of features. Boruta works by creating shadow features that mimic the original features' distribution of importance. It then compares the importance of the real features against those of the shadow features to determine their significance. Features that consistently exhibit higher importance than their corresponding shadow features are considered relevant and selected for further analysis. Boruta helps in improving accuracy and reducing over fitting by focusing on robust feature selection. A deatiled account of this method can be found in [13].

## 2.2 Prediction Algorithms

### Logistic regression

Logistic regression is a fundamental statistical method used in machine learning for binary classification tasks. Unlike linear regression, which predicts continuous outcomes, logistic regression models the probability that a given input belongs to one of two classes. It does this by applying the logistic function to the linear combination of input features, transforming the result into a probability score between 0 and 1. Logistic regression is widely used due to its simplicity, interpretability, and effectiveness in scenarios where understanding the relationship between independent variables and the probability of a specific outcome is crucial. It serves as a building block for more complex algorithms and is often the starting point for modeling classification problems.

### Decision trees

Decision trees recursively split the data based on the most informative features, creating a hierarchical structure resembling a tree. At each node, the algorithm selects the best feature to split the data, aiming to maximize information gain or minimize impurity. The process continues until the tree reaches a predefined stopping criterion. Decision trees are easy to visualize and understand, making them valuable for gaining insights into the decision-making process. However, they can suffer from overfitting if not pruned properly.

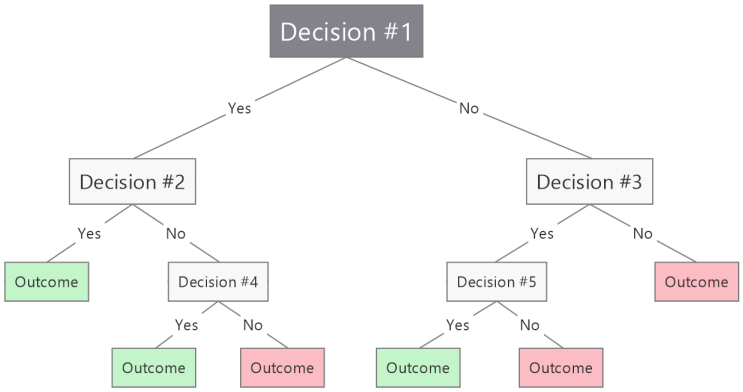


*Figure 2.1: Visual representation of decision tree [1]*

## Random forest

Random Forest is widely known for its robustness and accuracy in handling complex tasks. It is constructed by combining multiple decision trees, where each tree is trained on a random subset of the data and a random subset of features. During the training process, each tree independently makes predictions, and the final output is determined through a majority vote or averaging of the individual predictions. This technique helps mitigate overfitting and has the ability to handle large datasets and high-dimensional data.

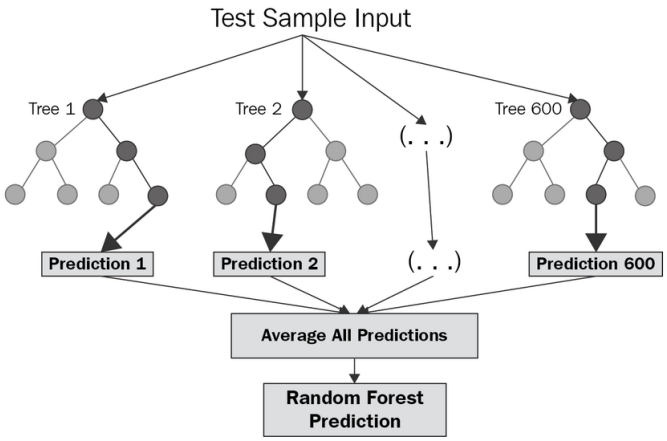


*Figure 2.2: Visual representation of random forest [14]*

## Neural networks

Neural networks are a type of machine learning algorithm that simulate the structure and function of the human brain. They are composed of interconnected nodes, or "neurons", that process and classify information. The input is transformed and processed through a series of hidden layers before producing the final output. Neural networks are trained through a process called backpropagation, where the weights and biases of the neurons are adjusted to minimize the error between the network's output and the desired output. The layered structure of neural network is shown in figure 2.3.

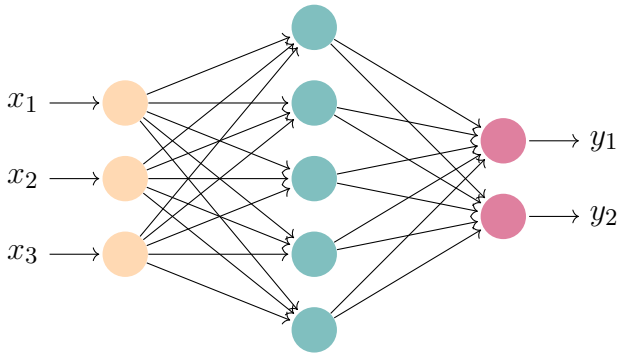


*Figure 2.3: Illustrates the layered architecture of ANN. Yellow nodes represent input variables, blue is a hidden layer with five nodes, and red represent output variables*

## Naive bayes

Naive Bayes is a simple yet effective probabilistic classification algorithm based on Bayes' theorem. It which calculates the probability of a specific class given the input features. The "naive" assumption in Naive Bayes is that all features are independent of each other, which simplifies the calculations and makes the model computationally efficient. Despite this oversimplified assumption, Naive Bayes often performs surprisingly well in practice, especially with high-dimensional and sparse data. Naive Bayes is quick to train and requires minimal tuning. However, its performance may be suboptimal when the independence assumption is significantly violated or when dealing with highly correlated features.

## Under-sampling techniques

Under-sampling techniques [11] are a set of strategies used to tackle class imbalance in machine learning datasets. When one class is significantly underrepresented compared to the other, models can become skewed and biased. Under-sampling involves reducing the instances of the majority class to create a more balanced distribution. While these techniques address class imbalance, they might lead to information loss from the majority class. Hence, careful consideration is required to strike the right balance between achieving class equilibrium and retaining the crucial data. Under-sampling methods play a vital role in achieving fairness and accuracy in predictive models dealing with imbalanced datasets.

# 3 Method

## 3.1 Literature Review

One crucial aspect to take into account is the choice of an appropriate method for generating predictions. For that purpose we conducted a traditional literature review to identify most suitable machine learning algorithm for our dataset. We searched for literature on websites like Google Scholar and ScienceDirect. We used simple words like classification, machine learning, ECMO, prediction, and machine learning ARDS to find the right articles. We could find only one article related to prediction of ECMO survival using machine learning [2]. In this article they utilised neural networks for prediction by analysing electronic records of 282 patients undergoing the therapy and achieved a prediction accuracy of 82%. To broaden our scope, we extended our search to include articles that analyzed datasets with similar attributes as ours, such as dimensions and other properties. In [21], 48 articles were used to compare several supervised machine learning algorithms for disease prediction. The article indicated that Support Vector Machine (SVM) was the most commonly employed algorithm, followed by Naive Bayes, while random forest demonstrated superior accuracy. A comparative study of five machine learning techniques (probabilistic neural network, perceptron-based neural network, random forest, one rule, and decision tree) is performed in [6] for predicting mesothelioma which is a type of lung cancer. The study analysed medical records of 324 patients and identified random forest as the best classifier. [5] analysed electronic medical records of 299 patients with heart failure for binary classification of survival. The classifiers included linear regression, Random Forests, One Rule, Decision Tree, Artificial Neural Network, two Support Vector Machines (linear and Gaussian radial kernel), k-Nearest Neighbors, Naive Bayes and gradient boosting. They found that random forest outperformed all the models.

After reviewing similar articles, we found that healthcare data often requires comparison between the algorithms to identify the suitable one. Therefore, we decided to use logistic regression, decision trees, random forest, neural networks, and Naive Bayes for our analysis. The reason for choosing these methods is discussed in the section 3.4.

## 3.2 Dataset

The dataset used for this study contain medical records of 450 patients. There were 145 women and 305 men among the patients, and their ages ranged from 17 to 83. All the patients had ARDS and 439 of them had undergone ECMO treatment.

The original dataset contain biomarker data of patients both before and after ECMO treatment. Since our aim is to classify patients based on their biomarker data before therapy, we retained only those columns that contain preECMO data. Additionally, three variables in the dataset were eliminated since they were irrelevant for this study. After performing these operations, our dataset contain 450 observations and 33 variables: 2 binary, one categorical and 30 continuous. A brief description of all the continuous variables is given in appendix 1, table A.1. Furthermore, we slightly changed the names of the features for the sake of clarity. Beyond the definitions of some of the features, no further details about the dataset or the features included were provided in the dataset manuscript.

## 3.3 Data preprocessing

Similar to [5], we employed the Shapiro-Wilks test [18] to assess the normality of each feature. We found that some of the features were normally distributed while others were not. Because of this, the correlations in the dataset were evaluated using both the Pearson correlation coefficient (parametric test) and the Spearman correlation coefficient (non parametric test). If the correlation between variables exceeded 0.7, we decided to retain only one of the variables to avoid problems when implementing prediction models. In our case, the variable with the highest number of missing values would be removed.

Twenty continuous variables out of the 33 variables in the dataset contain missing values (see fig3.2). The variable with the largest number of missing values is *albumin* (46%) followed by *ATIII* (6%) and *CRP* (6%). Since no explanation was provided about their absence, we decided to eliminate the variable *albumin* from the analysis and remove the rows containing NAs for *ATIII*. Following this procedure, the number of missing values for each patient is calculated. The maximum number of missing values per patient was 6. If the number of missing values in a row exceeded 3,

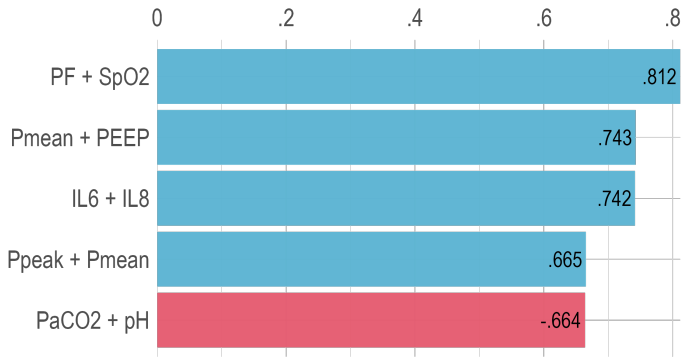


*Figure 3.1: Top five correlations in the dataset obtained using spearman correlation coefficient. Significance limit is 5%.*

we eliminated that row. Due to the relatively less number of NAs after this procedure, the remaining values were handled as follows:

- If the percentage of missing values is less than 2%, replace it with the median, and
- if the percentage of NAs is between 2-4%, replace it with random values from that column (instead of median to prevent bias).

The imputed dataset contain 29 features; 26 continuous, two binary and one categorical. The categorical variable in the dataset is a disease indicator and has seven levels (see fig. 3.3). We used one hot encoding technique to binarize this variable. One of the binary variables in the dataset is gender and the other is a ECMO survival indicator. The survival indicator, which indicates whether the patient survived the therapy or not, serves as the response variable in our binary classification analysis. Of the 450 patients, 341 (75.8%) survived the treatment and 109 (24.2%) did not survive the treatment. It is worth noticing that we have some data imbalance. The final dataset contain 403 observations and 35 features.

## 3.4 Data splitting and implementation

Feature selection methods like Lasso and Boruta uses a built-in cross-validation method. Therefore, the entire dataset was used when implementing these techniques. In the case of prediction models, we used two different strategies for data splitting. As decision trees, random forest and Naive Bayes does not require hyperparameter optimisation, we split the entire dataset in to training (80% of data instances randomly selected)

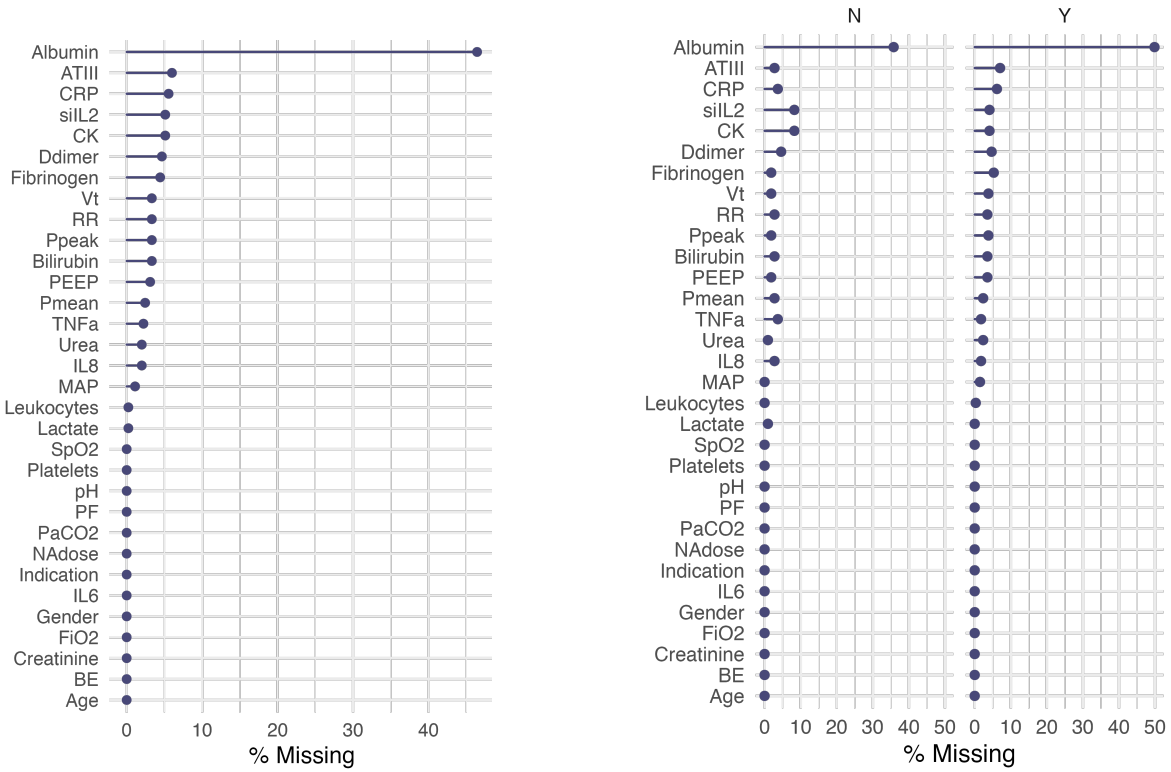


*Figure 3.2: Percentage of Missing Values in Each Variable in the Dataset: The first graph illustrates the percentage of missing values present in each variable across the entire dataset. The second graph displays the percentage of missing values in both the group of patients who survived ECMO ("Y") and the group who did not survive ("N").*

and test (remaining 20% ) sets. For neural networks, we need to do hyperparameter tuning. Therefore, we split the dataset in to training (60% of data instances randomly selected), validation (20% of data instances from the remaining) and test (remaining 20%) sets.

Machine Learning (ML) experts often recommend beginning with simple algorithms. So, we started with logistic regression utilising all the variables in the dataset. As logistic regression and Lasso are both regression-based models, we built an additional regression model using only the variables selected by Lasso. Another simple algorithm is the Naive Bayes classifier. Therefore, in addition to logistic regression, we also developed a Naive Bayes classifier. Given Lasso's ability to tackle problem of multicollinearity, we also constructed another Naive Bayes model using Lasso-selected variables.

Afterwards, we decided to move to tree-based algorithms because these methods are extensively utilised in biomedical contexts and are the least affected by mul-

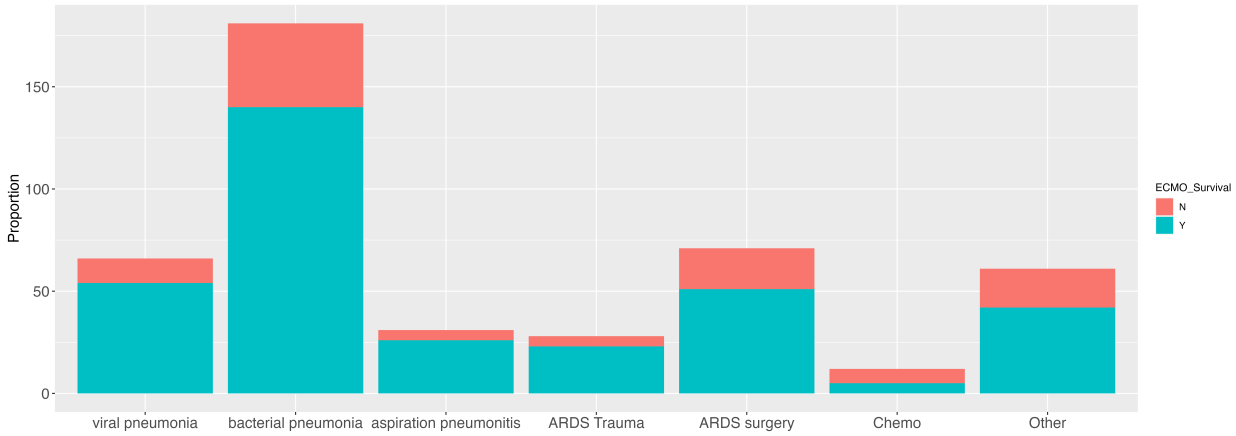


*Figure 3.3: Illustrates the proportion of different disease indicators in the dataset.*

ticollinearity in the data. We constructed decision trees and random forest models using all the variable in the dataset. Additionally, we designed an extra decision tree and random forest classifier using only the variables chosen by Boruta. This approach was taken because all these algorithms are tree-based.

Finally, following the previous literature we trained a neural network. Specifically, we developed four neural network models using the training dataset, each characterized by a distinct configuration of hidden layers and nodes: one layer and five nodes, one layer and 10 nodes, two layers with 10 nodes in the first layer and five nodes in the second, and two layers with 10 nodes in the first layer and 75 nodes in the second. These configurations were selected in accordance with previous literature. We then chose the model with the highest MCC value when applied to validation set and then generated prediction results through test set.

## 3.5 Performance metrics

The most popular performance metrics for binary classification problems are F1 score and accuracy. However, in cases where the dataset is imbalanced, this metric can be over optimistic. Matthews correlation coefficient (MCC, eqn.3.1) maybe a reliable metric in such situations as it produces a high score only if the prediction obtained good results in all of the four confusion matrix categories [4].

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{3.1}$$

Here TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives. MCC can take any value between -1 (worst value) and +1 (best value).

In this project, we used MCC for model optimisation and evaluation. To measure the effectiveness of our prediction models, we also used F1 score and accuracy.

# 4 Results

This chapter presents the results obtained using various algorithms. All the methods were implemented accordind to section 3.4 and evaluated according to section 3.5.

## 4.1 Feature selection results

We utilised lasso regression and boruta algorithm for feature selection. Lasso regression performs automatic variable selection by setting the coefficients of irrelevant variables to zero. Out of the 34 explanatory variables in our dataset, lasso selected five variables as important (see fig.4.1).

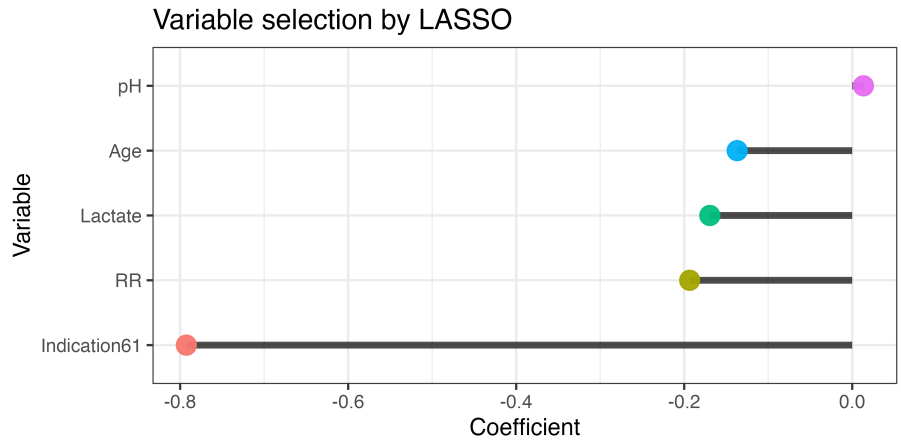


*Figure 4.1: Represents the variables selected by LASSO on y-axis and their coefficients on x-axis*

Besides lasso, we also used Boruta for feature selection because it can capture non-linear relationships and interactions in the data. Boruta identified eight variables as important in our dataset (see fig 4.2).

Both Lasso and Boruta algorithms identified Indication 6, Respiratory rate (RR), and Lactate as significant features. Lasso pinpointed five key variables, while Boruta confirmed six variables as important and marked two as tentative. It's worth noting that the variable selection by Boruta was inconsistent across different runs, with only Lactate, Indication 6, and respiratory rate consistently emerging as significant. The disparity in chosen variables between these two algorithms could stem from their

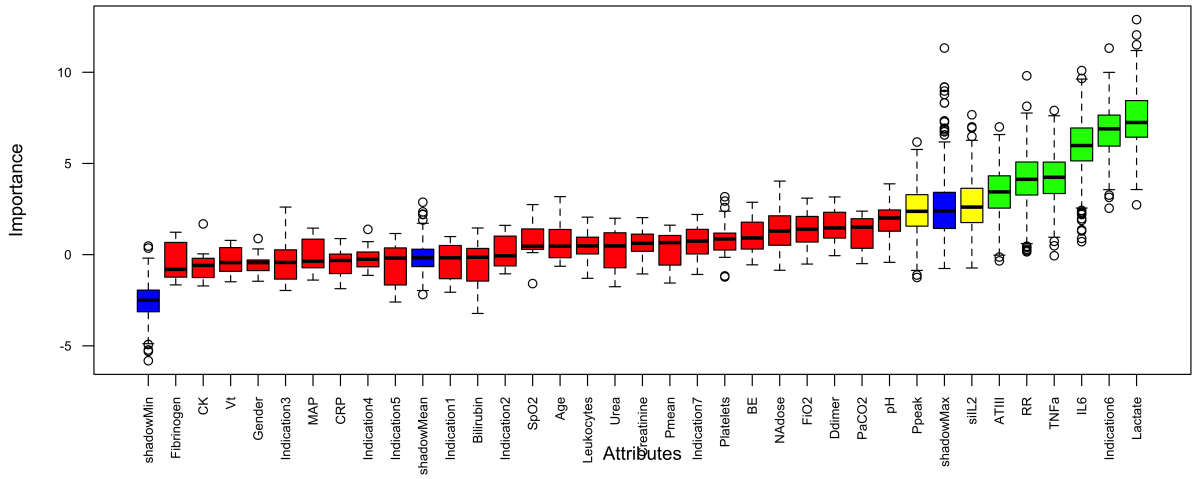


*Figure 4.2: Result of Boruta algorithm applied to entire dataset. Green box-plots represents confirmed important variables, red box-plots are confirmed unimportant variables, and yellow box-plots are tentative. Maximum number of importance source runs=500.*

distinct underlying statistical principles. In any case, both methods demonstrated a significant reduction in features, down to approximately 20%.

## 4.2 Prediction results

Following feature selection, we leveraged several ML algorithms for generating prediction results. We used logistic regression, decision trees, random forest, neural networks, and Naive Bayes. The MCC, F1 score and accuracy of all these models are given in table 4.1.

In general, all the methods achieved good accuracy. Random forest achieved high scores in all the three evaluation metrics while Naive Bayes (full model) attained lowest scores. Regarding regression models, logistic regression applied to Lasso-selected variables is the best classifier (MCC: 0.24), and performed better than the full model (MCC: -0.01). Similarly, Naive Bayes using Lasso-selected variables (MCC: 0.08) preformed slightly better than the full model (-0.08). This implies that utilizing only five variables is sufficient, rather than incorporating all variables present within the dataset. For tree-based models, random forest applied to variables selected by Boruta is the best classifier (MCC: 0.27) while decision trees applied to all the variables in the dataset is the worst classifier (MCC: 0). Similar to the regression models,

| Model | MCC | F1 | Accuracy |
|---|---|---|---|
| Logistic regression | -0.0091 | 0.1418 | 0.7160 |
| Logistic regression (Lasso) | 0.2457 | 0.2727 | 0.8025 |
| Decision trees | 0 | NaN | 0.7901 |
| Decision trees (Boruta) | 0.1752 | 0.2963 | 0.7654 |
| Random forest | 0.2169 | 0.1111 | 0.8024 |
| Random forest (Boruta) | 0.2731 | 0.3333 | 0.8024 |
| Naive Bayes | -0.0752 | 0.1290 | 0.6667 |
| Naive Bayes (Lasso) | 0.0857 | 0.1290 | 0.6667 |
| Neural networks | 0.1415 | 0.3721 | 0.6625 |

*Table 4.1: Results of ECMO survival prediction on complete imbalanced dataset; Dataset imbalance 300 survived (positive instances), 103 died (negative instances). Number of trees in random forest=500, Neural network learning rate=0.01.*

here also models with fewer variables performed better.

Generally, random forest built using Boruta-selected variables outperformed all the other models in terms of MCC, but was outperformed by neural networks in terms of F1 score and by logistic regression (with Lasso variables) in terms of accuracy. The MCC values of all the methods are close to zero, meaning these classifiers are no better than a random guess classifier. We believe that this is due to data imbalance and therefore we addressed this issue with the under-sampling technique. The implementation and results of the under-sampling technique is discussed in section 4.3.

## 4.3 Under sampling techniques

Even though we obtained generally good model accuracy on complete dataset, the MCC and F1 scores remained comparatively low. Given that MCC requires both high percentage of true positive and true negative predictions for a high score, this discrepancy is likely due to data imbalance. In other words, during the training process, each model becomes proficient at identifying positive instances (survived ECMO treatment), but struggles to accurately detect negative instances (didn't survive ECMO treatment).

Numerous techniques exist for managing data imbalance. An overview of these methods can be found in [12]. The issue of data imbalance in our dataset was tackled using the Random Under-Sampling (RUS) technique. RUS was chosen due to its

simplicity, speed, and the fact that it does not involve manipulation of data instances. To implement this technique, we created a balanced dataset such that it contain all the instances from the minority class and randomly selected instances from the majority class. The under-sampled dataset contain information of 101 patients who survived the treatment and 103 who did not survive the treatment. After that, we applied all prediction methods to this under-sampled dataset and the results obtained are given in table 4.2.

| Model | MCC | F1 | Accuracy |
|---|---|---|---|
| Logistic regression | -0.0996 | 0.3684 | 0.4146 |
| Logistic regression (Lasso) | 0.2587 | 0.5 | 0.5610 |
| Decision trees | -0.0644 | 0.5714 | 0.4878 |
| Decision trees (Boruta) | -0.0566 | 0.4102 | 0.4390 |
| Random forest | 0.1051 | 0.4 | 0.4878 |
| Random forest (Boruta) | 0.1924 | 0.55 | 0.5610 |
| Naive Bayes | 0.1459 | 0.5122 | 0.5652 |
| Naive Bayes (lasso) | 0.2981 | 0.6154 | 0.5122 |
| Neural networks | 0.7218 | 0.8148 | 0.8750 |

*Table 4.2: Results of ECMO survival prediction on balanced under-sampled dataset; Dataset imbalance 101 positive instances, 103 negative instances. Number of trees in random forest=500, Neural network learning rate=0.01.*

Compared to the previous results, here all methods achieved better F1 scores but lower accuracy, possibly stemming from information loss. Similar to previous results, the models applied to variables selected by feature selection algorithms performed slightly better than the full model. But, decision trees applied to Boruta-selected variables exhibited inferior performance in comparison to the full model. Neural networks outperformed all the other models in terms of all the three metrics (MCC: 0.72, F1: 0.81, accuracy: 0.88).

By integrating the outcomes of feature selection and prediction, it becomes evident that medical practitioners should direct their attention mainly to Lactate levels, chemotherapy status (Indication 6), and respiratory rate (RR). These key factors could offer valuable insights in determining the appropriateness of ECMO therapy for a patient.

# 5 Discussion

This project involves an in-depth analysis of suitability of ARDS patients for ECMO therapy. Our investigation is motivated by the critical need to identify optimal treatment pathways for patients with severe respiratory distress. By examining a range of patient biomedical markers, we seek to contribute to the medical community's understanding of the suitability of ECMO for acute respiratory failure, and inform evidence-based treatment strategies.

In order to identify the most important features (biomarkers) in our dataset, we first employed two feature selection algorithms. Both Lasso and Boruta consistently identified lactate, chemo (Indication 6), and respiratory rate (RR) as significant variables. The noteworthy role of lactate levels in ECMO treatment gains further validation from established research articles such as [7]. This suggests that doctors should consider the values of these biomarkers when assessing the suitability of ECMO therapy for a patient.

Following feature selection, our next step involved the construction of predictive models for ECMO survival. On the complete imbalanced dataset, logistic regression applied to Lasso-selected variables attained highest accuracy closely followed by random forest with variables selected by Boruta. However, random forest applied to Boruta-selected variables performed better than logistic regression based on MCC and F1 score. On the other hand, neural networks achieved the highest scores in all the three metrics (accuracy, F1 score and MCC) on a balanced under-sampled dataset. Based on these findings, for similar analyses, it's advisable to consider using random forest for imbalanced datasets and neural networks for datasets with a more balanced distribution.

The models built using the selected variables performed better than the full model, confirming the benefits of our feature selection. However, it's important to note that the MCC values of these models were closer to zero implying that the classifiers are very close to a random guess classifier. To further improve results, we plan to explore different variable selection methods in future analyses. This could provide a more complete picture of how variables interact and potentially lead to better predictions.

An important point to highlight is the substantial variability in these evaluation metrics across different runs. Notably, neural networks consistently maintained high MCC, F1 score, and accuracy on the under-sampled dataset, outperforming other models. It's advisable to consider the average of these metrics for a more comprehensive assessment of model effectiveness. In the future, we aim to utilize average metric values to enhance model evaluation and explore the reasons for this observed variability.

Although under-sampling addressed data imbalance, it came at the cost of losing valuable information. To enhance our approach, we also plan to explore different methods for handling data imbalance in future work, aiming to strike a better balance between preserving information and addressing the imbalance.

# A Dataset

| Feature | Range | Median | NAs |
|---|---|---|---|
| Respiratory Rate (RR) | [7,..,60] | 23 | 15 |
| Tidal volume (Vt) | [6,..,941] | 477 | 15 |
| Inspired fraction of oxygen (FiO2) | [0.21,..,1] | 1.00 | 0 |
| Peak airway pressure (Ppeak) | [15,..,50] | 34 | 15 |
| Mean airway pressure (Pmean) | [5,..,40] | 22 | 11 |
| Positive end expiratory pressure(PEEP) | [2,..,35] | 15 | 14 |
| Arterial partial pressure of oxygen(PF) | [28,..,409] | 69 | 0 |
| Periperal oxygen saturation (SpO2) | [29,..,99] | 90.95 | 0 |
| Arterial partial pressure of carbon dioxide (PaCO2) | [30,..,237] | 62 | 0 |
| Arterial pH (pH) | [6.39,..,7.57] | 7.23 | 0 |
| Arterial base excess (BE) | [-39,..,32] | -2 | 0 |
| Arterial lactate (Lactate) | [3,..,336] | 17 | 1 |
| Noradrenaline dose (NAdose) | [0,..,6.94] | 0.28 | 0 |
| Mean arterial pressure (MAP) | [34,..,109] | 68 | 5 |
| Creatinine | [0.1,..,11.6] | 1.25 | 0 |
| Urea | [2,..,703] | 58 | 9 |
| Creatinine Kinase (CK) | [9,..,36102] | 200 | 23 |
| Bilirubin | [0.1,..,29.6] | 0.8 | 15 |
| Albumin | [6,..,41] | 22 | 209 |
| C reactive protein (CRP) | [1,..,569] | 152 | 25 |
| Fibrinogen | [40,..,1236] | 510.5 | 20 |
| Ddimer | [1,..,36] | 6 | 21 |
| Anti-thrombin III (ATIII) | [10,..,650] | 63 | 27 |
| Leukocytes | [0,..,91.5] | 13.1 | 1 |
| Platelets | [2,..,808] | 182 | 0 |
| TNFa | [4,..,1468] | 25 | 10 |
| IL6 | [4,..,597450] | 461.5 | 0 |
| IL8 | [6,..,376513] | 113 | 9 |
| siIL2 | [27,..,121123] | 2183 | 23 |

*Table A.1: Statistical quantitative description of the continuous features in the dataset.*

# Bibliography


[1] L. Armstrong. Decision tree diagrams: what they are and how to use them. https://blog.mindmanager.com/decision-tree-diagrams/, 2021. [Online; accessed 03-August-2023].

[2] B. Ayers, K. Wood, I. Gosev, , and S. Prasad. Predicting survival after extracorporeal membrane oxygenation by using machine learning. *PLOS ONE*, 2020.

[3] G. Bellani, J. G. Laffey, T. Pham, E. Fan, L. Brochard, A. Esteban, L. Gattinoni, F. van Haren, A. Larsson, D. F. McAuley, M. Ranieri, G. Rubenfeld, B. T. Thompson, H. Wrigge, A. S. Slutsky, and A. Pesenti. Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries. *Journal of the American Medical Association. 2016 Feb 23;315(8):788-800*, PMID: 26903337, 2016.

[4] D. Chicco and G. Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics 21, 6 (2020)*, 2020.

[5] D. Chicco and G. Jurman. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med Inform Decis Mak 20, 16 (2020)*, 2020.

[6] D. Chicco and C. Rovelli. Computational prediction of diagnosis and feature selection on mesothelioma patient health records. *PLOS ONE*, 14(1):1–28, 01 2019.

[7] T. Datzmanncor and K. Träger. What about prognostic outcome parameters in patients with acute respiratory distress syndrome (ards) treated with veno-venous extracorporeal membrane oxygenation (vv-ecmo)? *PubMed*, 2018.

[8] F. Degenhardt, S. Seifert, and S. Szymczak. Evaluation of variable selection methods for random forests and omics data sets. *Briefings in Bioinformatics*, 20(2):492–503, 10 2017.

[9] M. Diamond, H. L. Peniston, D. K. Sanghavi, and S. Mahapatra. Acute respiratory distress syndrome. [updated 2023 apr 6]. *StatPearls [Internet]*, 2023.

[10] A. Esteban, F. Frutos-Vivar, A. Muriel, N. D. Ferguson, and et.al. Evolution of mortality over time in patients receiving mechanical ventilation. *Am J Respir Crit Care Med. 2013 Jul 15;188(2):220-30*, PMID: 23631814, 2013.

[11] H. A. et al. A review on data preprocessing methods for class imbalance problem. *International journal of Engineering and Technology*, 2019.

[12] K. M. Hasib, S. Iqbal, and e. a. Faisal Muhammad Shah. A survey of methods for managing the classification and solution of data imbalance problem. *Journal of Computer Science*, 2020.

[13] M. Kursa, A. Jankowski, and W. Rudnicki. Boruta - a system for feature selection. *Fundamenta Informaticae*, 2010.

[14] L. Lemeš and A. Akagic. *Prediction of Real Estate Market Prices with Regression Algorithms*, pages 401–411. 10 2022.

[15] K. Lewandowski. Extracorporeal membrane oxygenation for severe acute respiratory failure. *Epub*, 2000.

[16] G. Makdisi and I. wen Wangcorresponding. Extra corporeal membrane oxygenation (ecmo) review of a lifesaving technology. *J Thorac Dis.*, 2015.

[17] N. J. Meyer, L. Gattinoni, and C. S. Calfee. Acute respiratory distress syndrome. *The Lancet*, 2021.

[18] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[19] X. Teng, J. Wu, and S. X. Jing Liao. Advances in the use of ecmo in oncology patient. *Cancer medicine*, 2023.

[20] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[21] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis*

*Mak 19, 281 (2019)*, 2019.

[22] J. Villar, J. Blanco, J. M. Añón, and et.al. The alien study: incidence and outcome of acute respiratory distress syndrome in the era of lung protective ventilation. *Intensive Care Med. 2011 Dec;37(12):1932-41*, PMID: 21997128, 2011.