National College of Ireland

An Ensemble Learning Algorithm for ICU Patient Mortality Prediction

MSc Research Project
Data Analytics

Aoife Gaffney
Student ID: x19217781

School of Computing
National College of Ireland

Supervisor: Dr. Majid Latifi

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Aoife Gaffney
Student ID:	x19217781
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr. Majid Latifi
Submission Due Date:	23/09/2021
Project Title:	An Ensemble Learning Algorithm for ICU Patient Mortality Prediction
Word Count:	7634
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	23rd September 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

An Ensemble Learning Algorithm for ICU Patient Mortality Prediction

Aoife Gaffney
x19217781

Abstract

Prediction of patient mortality in Intensive Care Units (ICU) can aid the prevision of timely medical intervention and allocation of vital resources to those patients who are at the greatest risk of dying and for the provision of suitable interventions to save their lives. There is a lack of current research in an accurate, robust and timely solution that can handle complex imbalanced ICU data with significant missing values. This study examines the use of ensemble algorithms to produce reliable results in the prediction of ICU patient mortality by using patient's medical history data. Four popular single classifiers; Decision Trees (DT), Naive Bayes (NB), Logistic Regression (LR) and Support Vector Machines (SVM) and three ensemble classifiers; Light Gradient Boosting Machine (LGBM), Random Forest (RF) and Stacking are implemented in this research. Experiments are conducted with and without feature selection on a test set that handles data imbalance and missing values. The results indicate that the LGBM model without feature selection outperformed the state of the art approaches in terms of accuracy (0.97) and Area Under the Curve (AUC) (0.97). It was found that automated data-driven features selection did not improve the model performance if there was no prior domain knowledge.

Keywords- ICU, Patient Mortality, Classification, Stacking, LGBM, Ensemble Machine Learning, Feature Selection

1 Introduction

The application of data analytics to the medical domain has multiple benefits including reducing healthcare costs, improving overall healthcare practice and prediction of patient outcomes to improve medical care. The healthcare industry has vast amounts of data such as hospital records, data from machines and medical imagery (Raghupathi and Raghupathi; 2014). There has been an increased drive to digitalise this data to aid medical professionals in data driven decisions on patient care and overall public health management. This is particularly important in intense, time sensitive, clinical environments such as Intensive Care Units (ICU).

ICU are special units within hospitals that treat people who are critically ill therefore, timely medical intervention and effective resource allocation is critical. Due to this variety in ICU patient condition, specially trained healthcare professionals and monitoring equipment are required. ICU represent one of the largest clinical costs in hospitals. In the US, ICU costs currently account for almost 1% of the GDP, even though less than 10% of hospital beds are in ICU (Pirracchio et al.; 2015). Similarly, in the UK, the estimated cost of ICU per year is £541m which is 0.6% of NHS expenditure (Pirracchio et al.; 2015). The current strain on ICU due to the Covid-19 pandemic highlights the

importance of swift, data driven decisions in ICU. Covid-19 has brought under resourcing and over-crowding in ICU to the forefront. Due to the high cost and under-funding of ICU, it is necessary to triage resources (Kaier et al.; 2020). This involves assigning levels of urgency to patients in order to treat as many patients as possible. Predicting ICU mortality rates can aid this process and allows resources to be allocated to those patients who are at an increased risk of dying (Bhattacharya et al.; 2017).

Although research has previously been conducted in this domain, current clinical scoring systems are not sufficient to deal with complex ICU data which has unique patient conditions that are prone to outliers (Xu et al.; 2017). In instances of severely under-resourced ICU and unavoidable triages it is necessary to allocate beds or ventilators to those patients with a higher risk of mortality. There is a lack of application of ensemble methods to the domain, with the majority of studies focusing on single machine learning methods only. The majority of studies do not address the issue of imbalanced ICU datasets, with large amounts of missing data and inconsistent features. Although condition-specific models work well such as Rayan et al. (2021), there is a need for more general mortality prediction models as patients may suffer from multiple conditions leaving condition specific models redundant.

Considering all previous studies, this research focuses on applying seven different machine and ensemble learning methods both with and without feature selection for predicting ICU patient mortality based on patient's medical history. Decision Trees (DT), Naive Bayes (NB), Logistic Regression (LR) and Support Vector Machines (SVM) and three ensemble classifiers; Light Gradient Boosting Machine (LGBM), Random Forest (RF) and Stacking are examined to predict ICU patient mortality. The models are applied to a dataset with all features post preprocessing and a dataset consisting of a subset of the top 20 features selected by feature selection. This examines whether automated feature selection of the top 20 features improves model results. These features are then ranked by importance in predicting patient mortality.

The results of the final model can be used to aid in medical decision making leading to the reduction of unnecessary spending and inefficient resources in the ICU enables better allocation of funds elsewhere in the system. The use of a feature selection method will allow the most important features in the prediction of patient mortality to be available to medical staff which could act as an early warning system. This will enable medical staff to identify the high risk patients and act urgently to save lives, which is currently of critical importance due to the strain of Covid-19 pandemic on ICU resources. The research question on which this research is based is as follows:

"To what extent can ensemble methods deliver reliable results in the prediction of ICU patient mortality using patient's medical history data?"

The objectives of this project are as follow:

1. Critically evaluate various methods and models from relevant current literature.
2. Extensive pre-processing of data using suitable methods.
3. Creation of a subset dataset with top 20 features in prediction of ICU mortality using an automated feature selection method.
4. Ranking of the top 20 features in prediction of ICU mortality.
5. Implementation of single classifiers and ensemble classifiers on both datasets with and without feature selection.

6. Evaluation of implemented models with appropriate evaluation methods.
7. Comparison of models to other classifiers in current literature to assess performance.

Following this introduction, Section 2 covers an analysis of the current related work in the area of using machine learning for predicting ICU patient outcome, highlighting key findings and limitations of applied techniques. Section 3 presents the proposed research methodology. Section 4 describes the project design in detail. Section 5 addresses the implementation undertaken to answer the research question and objectives of this project. Section 6 contains a thorough evaluation of the results and main findings of the research. Finally, Section 8 concludes the research and discusses future work.

2 Related Work

The following section reviews and critiques recent relevant studies in the domain. The section begins with traditional scoring systems and moves onto machine learning and deep learning methods. Following this, ensemble methods and feature selection are discussed in detail. The section concludes with a comparison table of the most important research papers.

2.1 Current Scoring Systems

Earlier detection and diagnosis of ICU patient's outcome allows a greater possibility of saving lives. Over the past few decades, clinical scoring systems such the standard Acute Physiology and Chronic Health Evaluation (APACHE) set out by Zimmerman et al. (2006) and Sequential Organ Failure Assessment (SOFA) proposed by Ferreira et al. (2001) have been used to predict ICU mortality. These systems predict patient outcome and resource allocation using patient physiological and demographic parameters obtained in the first 24 hours after ICU admission (Awad et al.; 2017). These scoring systems are the current solution to assessing patient condition in ICU and several studies have been preformed indicating these methods preform adequately at predicting patient outcome in ICU with Area Under the Curve (AUC) of 0.84 achieved by Gilani et al. (2014) and AUC of 0.82 by Bennett et al. (2019). However, more recently, van Doorn et al. (2021) compares machine learning models versus clinical evaluation for mortality prediction in patients with sepsis using XGBoost and compares the results to standard ICU SOFA scores. The XGBoost significantly outpreformed the traditional SOFA method (Sensitivity = 0.92 vs Sensitivity = 0.72).

Although advancements have been made in the scoring systems over the years, they are not sufficient to deal with the complex ICU data with unique patient conditions (Xu et al.; 2017). There is a gap in the current research where prediction models are not accurate enough and there is no current tool in place to predict patient mortality in a timely and reliable manner (Kim et al.; 2011). These traditional scoring methods often use small condition specific datasets and are not adequately calibrated. Most severity scoring methods rely on LR models and these models put tight constraints on the relationship between the dependant variable and risk of mortality. It is suggested by Pirracchio et al. (2015) that this is unrealistic, considering the nature of ICU data, and may be the reason for low predictive power in some methods.

2.2 Machine Learning & Deep Learning Methods

In more recent years, there has been a shift towards using data mining and machine learning models to predict ICU patient outcome due to its ability to analyse large datasets

and identify patterns in the data. SVM, DT and LR are used by Lee et al. (2016) to predict ICU mortality with DT preforming poorest with an AUC of 0.6, which is only moderately better than random guessing. However, research by Kim et al. (2011) suggests that DT outperform non-linear models such as SVM with AUC = 0.89 and AUC = 0.87 respectively. Both models performed better than the APACHE III while using fewer variables for prediction, which also concurs with the findings of Pirracchio et al. (2015). In research carried by by Darabi et al. (2018) gradient boosted machines (GBM) outperformed Neural Networks (NN) with an AUC of 87% verses AUC of 77%.

More recently, RF outpreformed both GBM and LR in a study by Kong et al. (2020) to predict mortality of sepsis patients in the ICU with RF AUC = 0.85 verses LR AUC = 0.83. Similarly, RF outperformed NN and DT in mortality prediction in cerebral hemorrhage patients in ICU (Nie et al.; 2021). Naïve Bayes and Bayes Net are examined by Veith and Steele (2018) to predict ICU patient mortality using non-clinical attributes with AUC = 0.69 and AUC = 0.72. Alternative fuzzy models are examined by Silva et al. (2018) with 86% prediction accuracy and Hsieh et al. (2014) with AUC values of 0.85. In more recent years, Hou et al. (2021) and Ryan et al. (2020) examine the use of XGBoost for Covid-19 related ICU mortality prediction with good results of 89% AUC in Hou et al. (2021). Although machine learning methods yield good results, there is potential for improvement on these prediction models to ensure faster, more appropriate data driven decisions. The complex, continuously measured, time dependent data in ICU requires applying models that are more powerful and robust, without the need for additional dependent variables.

Deep learning models have become increasing popular in recent years due to the wide variety of GPU which allows faster, more powerful processing that has the ability to train itself unsupervised. Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) models are used by Zheng and Shi (2018). LSTM performs well against non-linear methods such as SVM and RF with LSTM achieving AUC of 0.80 verses RF AUC of 0.64 and SVM AUC of 0.56. Convolutional Neural Networks (CNN) are applied by Kim et al. (2019) for mortality prediction in critically ill children in ICU. CNN algorithm achieved the highest AUC in the range 0.89–0.97, indicating that it outperforms the current standard PIM 3 (Pediatric Index of Mortality 3) used in Paediatric ICU. However, it must be highlighted that this study is based on a small population estimate. A paediatric dataset is also examined by Che et al. (2016) who looks at Deep Neural Network (DNN) and RNN compared to LR and SVM. The best model was a combination of both DNN and RNN which produced AUC score of 0.78. A DNN is also used by Ahmed et al. (2020) to predict mortality in trauma patients admitted to ICU.

In contrast to Che et al. (2016) the research of Xu et al. (2017) focuses on combining sequential and non-sequential factors. Similar to Zheng and Shi (2018), the model is a RNN-LSTM based prediction model that outperforms LR with an average AUC of 0.80 verses AUC of 0.78. An issue in current methods is the lack of generalisation of results as many of the studies are conducted in controlled environments, so the effectiveness of the models often drops when the models are tested with different patient populations and conditions. This concern is addressed by Alves et al. (2019) who explores domain adaptation methods to improve the accuracy of prediction systems by using CNN and LSTM to extract shared latent features from ICU domains or patient sub-populations. The proposed model significantly outperformed the baselines including SVM and RF with AUCs = 0.9, AUC = 0.6 and AUC = 0.6 respectively.

Such conflicting results on the performance of different prediction tools reveal that

no single algorithm invariably outperforms all others; it depends on the population of interest, the variables measured and the outcome being tested. The lack of improvement in the performance of deep learning models over machine learning model indicates the need for a more comprehensive model.

2.3 Ensemble Methods

Although single machine learning and deep learning algorithms have produced good results, ensemble methods can be used to improve model performance (Ghorbani et al.; 2020). An ensemble method is a technique which uses multiple independent models or weak learners to derive an output and make more accurate predictions. Prediction of ICU mortality in the first 6 hours after admission is examined by Awad et al. (2017). An ensemble model of the best preforming models amongst RF, DT, NB and Projective Adaptive Resonance Theory (PART) is examined. The RF based model performed the best with AUC of 0.83, which outperformed APACHE scoring system (AUC of 0.65). A RF based ensemble method was also used by Rayan et al. (2021) for early sepsis prediction in ICU with 98% accuracy. More recently, El-Rashidy et al. (2020) applied a patient specific stacking ensemble model to predict ICU mortality. This method is more suitable than traditional methods as it can handle complex problems involving decision boundaries that lie outside the space of the function. Therefore, this reduces the chance of poor classifier selection. Ensemble methods also handle the complex and diverse data recorded in ICU better. The ensemble method out-performed single ML techniques such as KNN, DT and LR by 7-19% in terms of accuracy when computed on time series data, 6 hrs to 24 hrs after admission.

Another ensemble method is used by Ghorbani et al. (2020) where the model is based off Genetic Algorithm (GA) for feature selection with stacking and boosting to compute an early ICU patient mortality prediction model. The proposed model is a combination of Multilayer Perceptron Neural Network (MLP), KNN and Extra Tree Classier as base classifiers and Boosted SVM is selected as the meta-classier. This new model results in an average prediction accuracy of 81% in comparison to 77% for RF and 64% for DT (Ghorbani et al.; 2020). Similarly a RF ensemble method is used by Ghose et al. (2015) with prediction accuracy of up to 87%. More recently, a dynamic ensemble learning algorithm (DELAK) based on K-means is proposed by Guo et al. (2021) with excellent results of 87% for prediction of mortality within 72 hours. The issues of class imbalance, missing values and feature selection are discussed in detail as challenges by (Ghorbani et al.; 2020). It is explained by Awad et al. (2017) that overcoming these issues can lead to improved results for classifiers. Several solutions are discussed including Synthetic Minority Oversampling Technique (SMOTE) for oversampling. Several different techniques are proposed for feature selection and ranking including GA, LR, WEKA, or simply choosing those attributes with high availability/coverage, meaning that the attribute/test should be measured at least once for each patient Awad et al. (2017). It is also suggested to use the expertise of ICU consultants or attributes used in previous research El-Rashidy et al. (2020) and Kim et al. (2019). The majority of studies use the a combination of coverage and domain knowledge for feature selection including Bhattacharya et al. (2017) and Silva et al. (2018). The ranking of features is examined by Ryan et al. (2020) and Veith and Steele (2018) using F score and WEKA respectively to rank the most important features relative to the response variable. The importance of ICU mortality risk factors, especially in the current Covid-19 pandemic is highlighted in research by Monteiro et al. (2020) and Sanaie et al. (2021).

From the review of current literature, it is clear there is a lack of use of ensemble models in current literature which can handle the large volume incomplete nature of ICU data well. Most literature is based on single condition specific classifiers that do not handle the volatility and abundant ICU data. Some studies focus on one algorithm only or are condition specific and there is a lack of comprehensive evaluation methods in some studies such as Awad et al. (2017) and Rayan et al. (2021). Due to this, an ensemble method is proposed in this study alongside four single classifiers for comparison purposes with and without feature selection. The models are applied to a preprocessed dataset with feature selection that ranks the top 20 high risk features in ICU patient mortality and a preprocessed dataset without feature selection.

A summary of the most important relevant research studies for this paper is provided in Table 1. The following section will outline the research methodology adhered to in this project to explore the prediction of ICU patient mortality.

Table 1: Summary of Related Works

Author	**Objectives**	**Method**	**Advantages**	**Limitations**
Awad et al. (2017)	Early ICU mortality prediction 6 hours post admission	- SMOTE for oversampling - Random forest ensemble method	Proposed model outpreformed 3 different models with AUC = 0.89	Only one evaluation method is used (AUC)
Ghorbani et al. (2020)	Early ICU mortality risk 24 hrs post admission using imbalanced dataset	- Genetic Algorithm (GA) in feature selection - Stacking and Boosting ensemble method	Proposed model achieved average accuracy = 81% compared to 9 different models	Small dataset
El-Rashidy et al. (2020)	Model for early ICU mortality prediction 6 - 24 hrs post admission	- Stacking ensemble model based on 5 classifiers - Feature selection methods computed by ICU domain expert	Ensemble model preformed the best with AUC = 0.93	Requires significant computational power
Ghose et al. (2015)	Patient specific ICU mortality risk 48 hrs post admission	Random forest based ensemble classifier	Model achieved results of 0.87 accuracy	- Using only one evaluation method - Does not address class imbalance or feature selection.
Guo et al. (2021)	Dynamic prediction model for ICU mortality	Ensemble algorithm based on K-means sampling and distance-based dynamic ensemble model	Ensemble model preformed the best (AUC = 87 72 hrs after ICU)	- Searching for base classifier is time-consuming process.
Rayan et al. (2021)	Early sepsis prediction in ICU using clinical records	- Random forest based ensemble classifier - Undersampling and oversampling for class imbalance	Accuracy of 98% when compared to 10 classifiers	Single condition study on sepsis

3 Research Methodology

The goal of this study is to develop an ensemble machine learning model that can predict ICU patient mortality effectively and efficiently. The methodology followed for this research is based upon the CRISP-DM (Cross Industry Standard Process for Data Mining) framework as illustrated in Figure 1.

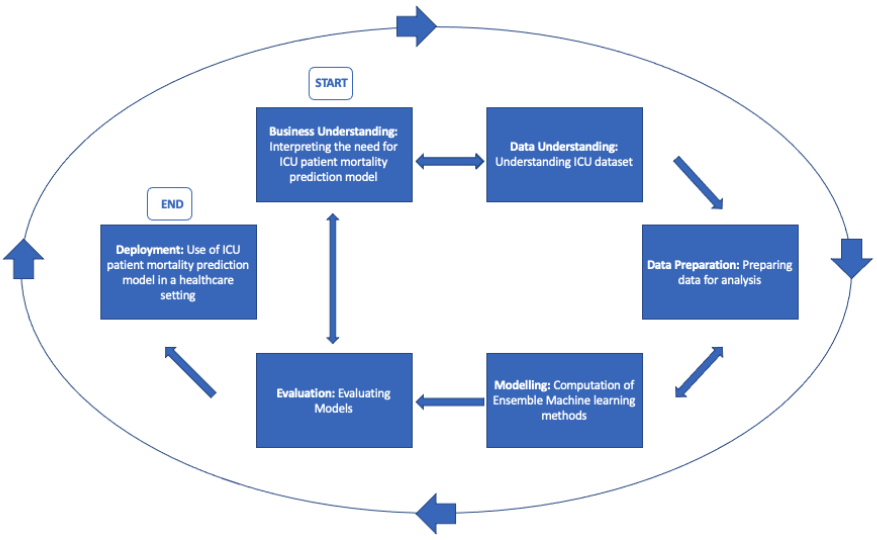

Figure 1: Research Methodology

3.1 Business Understanding

The project goal is to design an accurate and reliable prediction model to predict patient mortality in ICU. This will help to make data driven decisions to improve patient outcomes and allocate resources swiftly and effectively for the benefit of all patients. The project plan is to use several different types of ensemble and machine learning methods to achieve the project goal.

3.2 Data Understanding

The data in this project is taken from Kaggle using their open access WiDS (Women in Data Science) Datathon 2020: ICU Mortality Prediction competition dataset [1]. This is a large, freely-available database from MIT's GOSSIS community initiative. The dataset is comprehensive and comprises of deidentified hospital ICU visits from patients, spanning a one-year timeframe from hospitals worldwide. The data is taken from the first 24 hours post ICU patient admission to predict patient mortality. Data is collected following standard data acquisition protocols with consent from the patient regarding the use of data for the purpose of research work. The dataset contains 91,713 rows and there are 168 features that are divided into five main groups; identifiers, demographics, vitals, labs and APACHE values.

3.3 Data Preparation

The data must go through extensive pre-processing in order for it to be modelled correctly. This sections includes Exploratory Data Analysis (EDA), feature engineering,

[1] https://www.kaggle.com/c/widsdatathon2020

handling missing values, label encoding, handling class imbalance, feature scaling and feature selection as shown in Figure 2

Figure 2: Data Preparation

3.3.1 Exploratory Data Analysis (EDA)

EDA is performed to get an overall sense of the data and its underlying structure which is vital due to the nature of ICU data (Bhattacharya et al.; 2017). This includes a search for variation, missing values, outliers, skewness and correlation. Statistical (mode, mean, max, min) and graphical methods (boxplots, histograms) are computed.

3.3.2 Feature Engineering

Feature engineering refers to cleaning up or removing raw data to improve performance and is a key step in the process. This includes cleaning up features and removing the following;

1. Features with a high percentage of missing values (60% and up).
2. Collinear (highly correlated) features - with threshold above 0.99.
3. Features with zero standard deviation.
4. Features with all unique values.
5. Features with zero importance and zero influence.
6. Aggressive and highly important features.

3.3.3 Handling Missing Values and Outliers

Due to the nature of large ICU datasets, a check for missing values is necessary as missing values can lead to inaccurate results. Missing values and outliers are identified and any outliers or features containing over 60% missing values are removed from the dataset. All other missing values are compensated for using Multivariate Imputation by Chained Equation (MICE). This type of imputation works by filling the missing data multiple times in order to measure the uncertainty of the missing values accurately and provides unbiased estimates[2].

3.3.4 Category Encoding

As the dataset consists of various categorical variables including 'icu_admit_source' and 'gender', categorical encoding is necessary to ensure features are in the correct format for modelling. One Hot encoding is used in this project as it is simple and the result is binary rather than ordinal. It involves replacing the categorical values with a numeric value between 0 and the number of classes minus 1 as most machine learning algorithms will not understand categorical variables.

[2]https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779

3.3.5 Class Imbalance

Class imbalance is when the classes in a dataset are not equally represented and causes bias which is a common issue in clinical data. The data in this study are significantly imbalanced as only 8.6% of patients died in ICU. Oversampling methods such as SMOTE are used by Ghorbani et al. (2020) and Awad et al. (2017) and work by determining distances for the minority class samples near the decision boundary and creates new samples. Oversampling is preferred to undersampling as it doesn't lose any knowledge from the majority class and therefore the SMOTE method is chosen for this research.

3.3.6 Feature Scaling

Features scaling is an important step in data pre-processing in order to bring all features in the same standing so that one significant number does not impact the models. The dataset contain highly diverse features in various sizes, units, and ranges. Standardisation is used as utilised by Ghorbani et al. (2020) and all features are re-scaled so all of them have the characteristics of a standard normal distribution. Standardisation helps to decrease processing time and allow more accurate results.

3.3.7 Feature Selection

Feature selection is an important part of the process which involves selecting an optimal subset of relevant features to be used in the development of predictive models. This helps curtail the dimensionality of the dataset by ignoring the insignificant or noisy features and reduces risk of overfitting of data. This research uses a univariate feature selector that is simple to run and faster than wrapper methods on a large dataset. The feature selection tool used is GenericUnivariateSelect from sklearn tool that allows selection of features from a dataset using a scoring function. It works by selecting the best features based on univariate statistical tests and removes all but the k highest scoring features. The scoring function used is mutual information (MI). MI between two random variables measures the dependency between the variables. It is equal to zero if two random variables are independent, and a higher value indicates higher dependency between variables. It is a non-parametric test which uses k-nearest neighbors to measure the scale of the relationship between the predictor variable and the target variable. This allows it to select the 20 top features in the prediction of ICU mortality. Some advantages of this feature selection method is that is it an automated method with pre-defined statistical tests, easy to use and implement. A separate dataset containing the top 20 features in the prediction of ICU patient mortality is constructed to examine if implementing feature selection improves model results.

3.4 Modelling

From Section 2, this project focuses on the application of ensemble and machine learning approaches to achieve the research objective in Section 1. There are 3 mains types of ensemble learning; boosting, bagging and stacking. A model of each type of ensemble method (LGBM, RF and Stacking) will be applied in order to get the best preforming ensemble classifier along with 4 single classifiers required for the stacking model and for comparison purposes. The dataset is split into training and test set. Hyperparameter optimisation is computed by using 10-fold cross validation on the training set to avoid overfitting.

3.4.1 Single Classifier Models

SVM SVM is a popular supervised learning algorithm that is chosen because it can handle high dimensional data well. However SVM can under-perform in large, noisy datasets. SVM works by creating a hyperplane which separates the data into classes and is applied successful by Ghorbani et al. (2020) and by Rayan et al. (2021).

DT DT is a supervised learning algorithm that is popular due to it's intuitive nature and the ability to handle missing values well. However, DT can tend towards overfitting and works by using multiple algorithms to split a node into two or more sub-nodes. DT is selected as it produced good results in studies conducted by El-Rashidy et al. (2020) and Awad et al. (2017).

LR LR is a supervised learning classification algorithm that is easy and quick to implement but is not as powerful as the other algorithms and assumes linearity between variables. LR works by estimating the relationship between the target variable and predictor variables. LR achieves acceptable results in research conducted by El-Rashidy et al. (2020) and Ghorbani et al. (2020).

NB NB is a supervised learning classification model that is simple and suitable for large datasets. It works on the assumption that the presence of a particular feature in a class is unrelated to the presence of any other feature and separates data into different classes according to the Bayes' Theorem. NB is chosen due to simplistic nature, speed and effective handling of large data when utilised by Awad et al. (2017).

3.4.2 Ensemble Models

RF RF is a bagging method chosen due to its ability to handle large datasets and avoid overfitting as used by several studies including studies by Rayan et al. (2021) and by Ghose et al. (2015). RF is based on many DT and the applied model fits DT classifiers on different subsets of the dataset where a random subset of input features is used. Each tree in the ensemble is built from a sample drawn with replacement from the training data. The model aggregates the results from trained DTs through averaging.

LGBM LGBM is a bagging method which is a novel approach not used in any of the reviewed literature. LGBM is a relatively new algorithm allows fast training speed and handling of large scale datasets[3]. LGBM is a gradient boosting framework that uses tree based learning algorithms. It works by growing vertically leaf-wise and chooses the leaf with large loss to grow. LGBM was chosen as it uses less memory to run and can handle large amounts of data[4]

Stacking Stacking works by using several different base learners that are aggregated using another model (meta learner) which combines the decisions of the base learners as shown in Figure 4. Stacking is successfully implemented by El-Rashidy et al. (2020) and Ghorbani et al. (2020). The stacking method is used combine all classifiers to obtain a stronger ensemble classifier.

[3]https://lightgbm.readthedocs.io/en/latest/

[4]https://www.analyticssteps.com/blogs/what-light-gbm-algorithm-how-use-it

3.5 Evaluation

This step involves evaluating the model in line with the research objectives in Section 1. In order to evaluate the models accordingly, several metrics are chosen. Following a review of the literature in Section 2, the evaluation techniques selected are:

- **Accuracy**: Percentage of number of correct predictions to the total number of instances (Rayan et al.; 2021).
- **AUC**: How much the model is capable of distinguishing between classes (Ghose et al.; 2015)
- **Recall**: Percentage of positive instances out of the total actual positive instances (Ghorbani et al.; 2020)
- **Precision**: Percentage of positive instances out of the total predicted positive instances (El-Rashidy et al.; 2020)
- **F1 score**: The balance between the precision and the recall (El-Rashidy et al.; 2020)

4 Design Specification

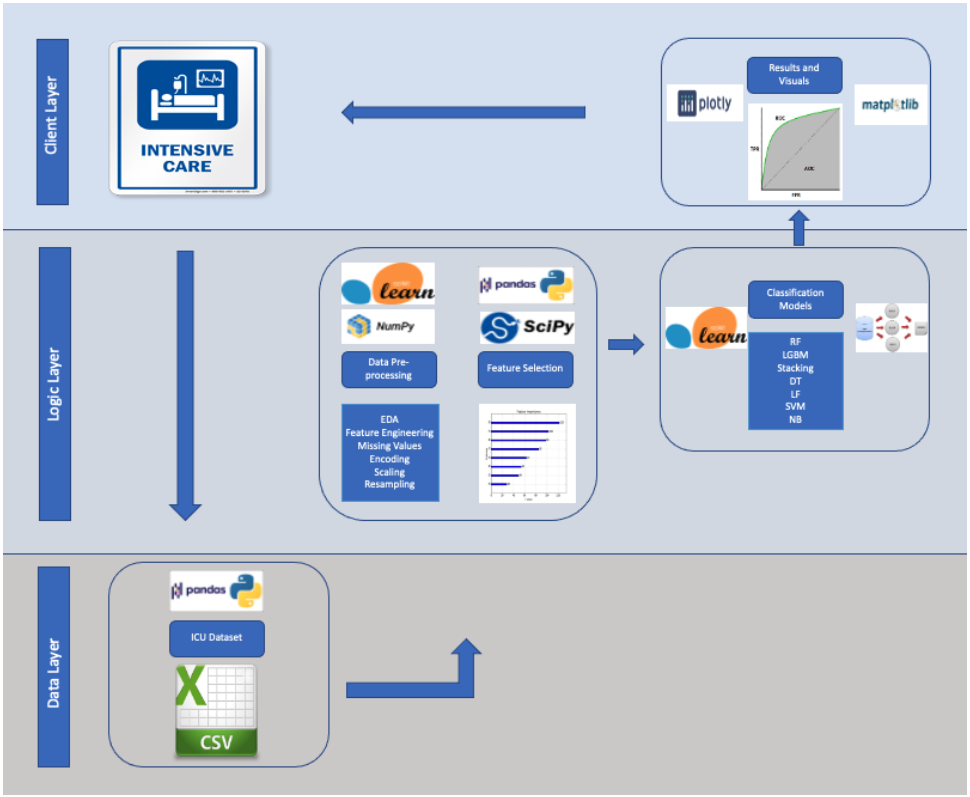

Figure 3: Project Design Specification

All practical experiments are carried out with a 1.6GHz dual-core Intel Core i5 MacBook Air with 8GB of RAM. The code runs on Google Colab and is coded in Python. Multiple python packages were imported for the execution of the project using pip and import functions. As mentioned in Section 3, the research is based on a CRISP-DM approach and the three tier design specification of the proposed research is shown in Figure 3. The process commences in the client layer with the gathering of the ICU dataset

and moves into the data layer where the data is downloaded in .csv format in Microsoft Excel, uploaded to Google Colab and transformed into a pandas dataset for analysing. The process moves into the logic layer where dataset analysed, pre-processed and modelled in Python with the skearn, scipy, pandas and numpy packages. Next, the process moves back to the client layer where model is evaluated, and results are produced and visualised with Matplotlib and plotly packages. The final stage is the deployment of the final classification model into the ICU environment. The following section examines the implementation of this research design.

5 Implementation

This section contains all steps in the implementation of this project including data preparation, modelling and evaluation.

5.1 Data Preparation

This section discusses the data set up, preparation and pre-processing before modelling. The necessary packages were loaded and the ICU dataset was downloaded in .csv format from Kaggle[5] and loaded into Google Colab as a pandas dataframe. Basic statistical information for each variable is computed with describe() and several plots are prepared to identify outliers, skewness and distribution of the data using the plotly and seaborn packages. Four features including 'icu_type' underwent data cleaning to clean up groupings within each feature as identified by the visual plots. As BMI had a high volume of missing values, it was computed manually using the formula: BMI = Weight(kg)/(Height(m)*(Height(m)). Eight redundant features were removed such as 'patient_id' and 'hospital_id' as they can decrease generalization performance and do not contribute to the analysis. These dropped features include those with zero standard deviation, those features with all unique values and features with zero importance. A check for correlation between non-categorical variables is undertaken to identify those variables with high correlation with the corr() function. Fifty-nine features with a correlation of 0.9 or higher are dropped in order to avoid skewed or misleading results.

A check for all missing data is complete with train.isna() function and 38 variables with over 60% missing values in the train set are dropped. The remaining missing values are compensated for using MICE. MICE is implemented on the numerical data with the SimpleImputer() function from sklearn and applied using 'mean' imputation value as it most suited to numerical data. The 'most frequent' strategy is used as the imputation value on categorical features which is suitable for text data. One Hot encoding is computed using panda's pd.get_dummies() for categorical data. All non-categorical features are standardised by using the sklearn prepossessing function StandardScaler() which removes the mean and scaling to unit variance. SMOTE is used to upsample the minority class (death) to the same number of samples as the majority class (survival) using the resample() function from sklearn.

GenericUnivariateSelect() feature selector from sklearn is applied to the dataset create a subset of top 20 features using a scoring function and a new pandas dataframe is constructed. The top 20 features are then ranked by importance in prediction of patient mortality and plotted using plotly package. The two datasets, the original and the dataset containing the top 20 features only, are then spilt into 70% train and 30% test splits for modelling.

[5]https://www.kaggle.com/c/widsdatathon2020/data

5.2 Modelling

This section discusses the steps for modelling and tuning of various classification models. In this study, 7 different classification models are examined. The performance of each model is evaluated to identify the best fitting model for classification. All training and tuning was conducted on the training dataset using 10-folds cross validation.

5.2.1 Modelling Without Feature Selection

First, DT, SVM, RF and LGBM base models are computed using the appropriate sklearn package functions on the original prepared dataset containing all features. Their optimised parameters are found using Randomized Search function RandomizedSearchCV() from sklearn.model_selection with 10-fold cross validation to prevent overfitting. Each model has it's own specified parameter search based on research from Section 2. NB is excluded from this step as it does not have hyperparemters to tune. The Randomized Search method is chosen due to it's decreased computation time in comparison to Grid Search. After research from Section 2, the search parameters for each model were decided upon based on a trade off between the optimal hyper parameters for each model and keeping the computational time as low as possible. Similarly, Linear SVM is utilised instead of SVM due to computational time constraints.

Next, the model is constructed of NB and all 3 of the single classifiers with their optimised parameters as detailed in Figure 4. As sklearn did not have an appropriate function, a new () function from vecstack package is implemented using mode='oof_pred_bag' and metric = area under the curve with 10-folds cross validation. OOF means out-of-fold prediction and refers to predictions for section of train data that the model hasn't seen during training. After computation, DT is selected as the best base classifier based on AUC and is applied to the dataset to get the final prediction.

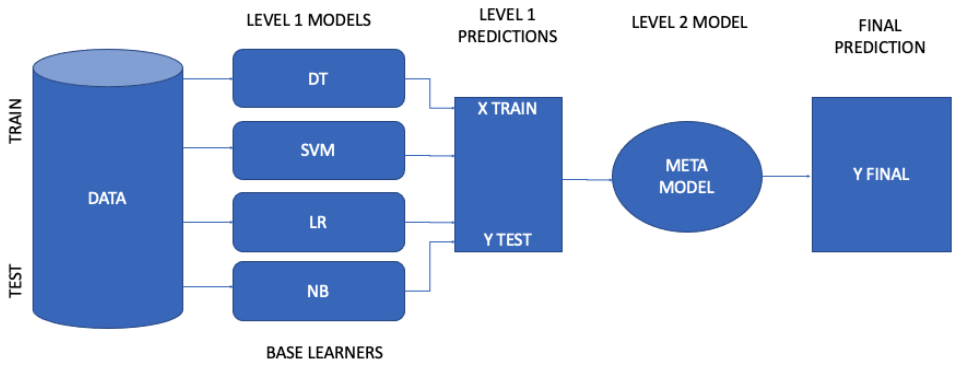

Figure 4: Stacking Methodology

5.2.2 Modelling With Feature Selection

This method is the same as Section 5.2.1. All models are applied to the subsection of the dataset containing the top 20 features identified by the GenericUnivariateSelect() function. As previously stated, the models are computed using the appropriate sklearn package functions. Another Randomized Search is conducted on each model due to the change in dataset with the function RandomizedSearchCV(). A model of the four single classifiers is constructed with vecstack package. As DT is also the best base classifier, it is applied to the dataset to get the final classifier prediction, Figure 4.

6 Evaluation & Results

This section details the results and evaluation of the classification models which relates to objective 6 in Section 1. The following evaluation methods are used to assess the performance of each model; AUC, Accuracy, Recall, Precision and F1 score. The evaluation methods are applied to each model with and without feature selection. As this is a classification study, the AUC evaluation metric is most frequently used in research from Section 2 and as the class imbalance has been address in this research, accuracy can also be relied upon as an evaluation metric.

6.1 Results Without Feature Selection

Table 2 shows that both LGBM and stacking models preformed the best with AUC = 0.97 and AUC = 0.97 respectively as seen in Figure 5. Both NB and LR preformed poorly with AUC = 0.76 and AUC = 0.72 respectively. However, LR does have higher accuracy (0.80), recall (0.79) and precision (0.80). RF and DT both achieved good results with DT AUC = 0.93 and RF AUC = 0.95. SVM achieved adequate results of AUC = 0.80. NB has a poor accuracy score indicating it is not very accurate at correct predictions. In terms of precision, recall and F1 score, LGBM and stacking both achieved excellent results of recall = 0.99, precision = 0.94 and F1 score = 0.97. This indicates they are able to correctly identify true positives (predicted as death and was death).

Table 2: Results Without Feature Selection

Model	AUC	Accuracy	Recall	Precision	F1 Score
NB	0.76	0.76	0.74	0.77	0.75
DT	0.93	0.93	0.97	0.89	0.93
SVM	0.80	0.80	0.78	0.80	0.79
LR	0.72	0.80	0.79	0.80	0.79
RT	0.95	0.95	0.97	0.93	0.95
LGBM	0.97	0.97	0.99	0.94	0.97
Stacking	0.97	0.97	0.99	0.94	0.97

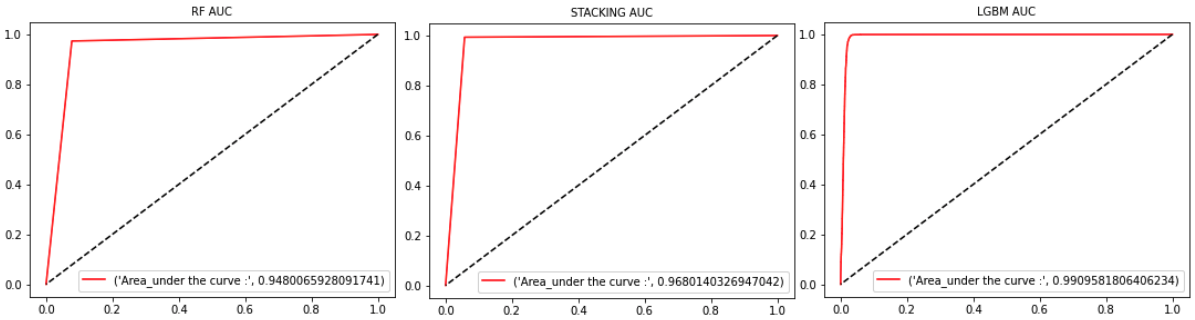

Figure 5: ROC AUC Curve for Ensemble Methods Without Feature Selection

6.2 Results With Feature Selection

In this experiment, the role of the feature selection to improve results is examined. Table 3 shows that similarly to Table 2, the LBGM preformed the best with AUC = 0.95

and accuracy = 0.95 respectively. In contrast, the Stacking model did not preform as well on this test set with AUC = 0.86, as seen in Figure 6. Both NB and LR preformed poorly, with both classifiers achieving AUC = 0.73 . However, LR does have higher accuracy (0.75) and precision (0.77). RF achieved very good results with AUC = 0.93 however, DT slightly underpreformed at AUC = 0.80 along side SVM with AUC = 0.75. LGBM and stacking both achieved excellent results of recall = 0.99, precision = 0.91 and F1 score = 0.95, which is only slightly lower than Table 2.

Table 3: Results With Feature Selection

Model	AUC	Accuracy	Recall	Precision	F1 Score
NB	0.73	0.73	0.75	0.72	0.73
DT	0.80	0.80	0.81	0.80	0.80
SVM	0.75	0.75	0.67	0.78	0.73
LR	0.73	0.75	0.73	0.77	0.74
RT	0.93	0.93	0.96	0.90	0.93
LGBM	0.95	0.95	0.99	0.91	0.95
Stacking	0.81	0.81	0.81	0.81	0.81

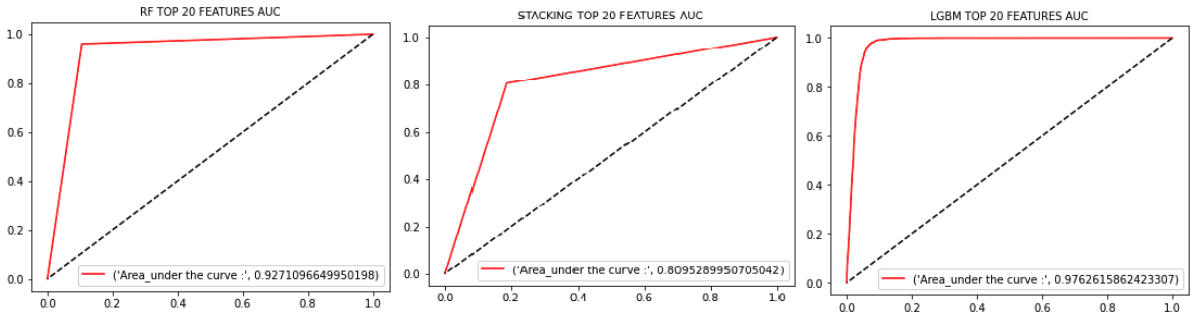

Figure 6: ROC AUC Curve for Ensemble Methods With Feature Selection

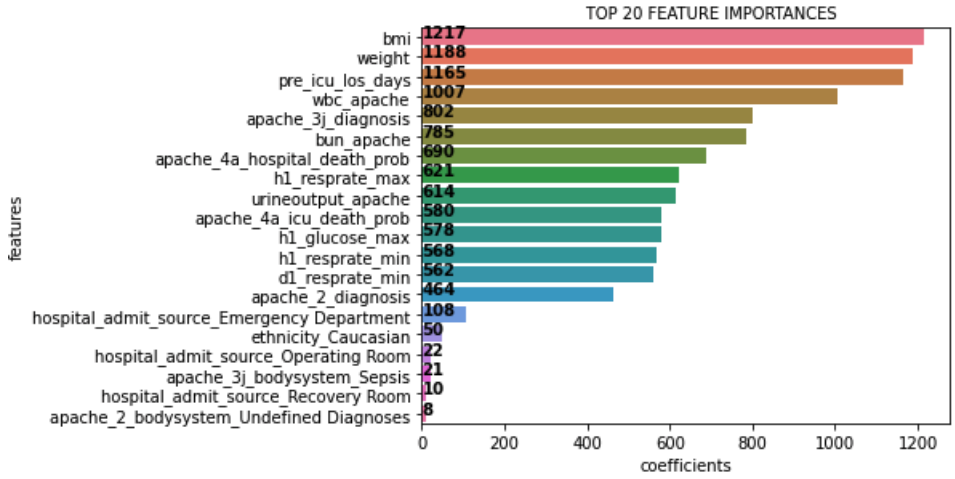

Figure 7: Top 20 Feature Importance

Table 2 results show that the models computed in the full dataset, without feature selection, outpreformed the models computed with feature selection, in Table 3. Overall, the LGBM model without feature selection preformed the best with AUC = 0.97. The poorest model performance is NB with feature selection with AUC = 0.73. The top 20 feature importance ranking is shown in Figure 7. BMI, weight and the length of patient stay prior to ICU admission are the top 3 features in prediction of ICU patient mortality. along with several APACHE and vital features.

7 Discussion

The aim of this project is to predict ICU patient mortality by applying ensemble machine learning methods to patients medical and demographic data as detailed in the objectives of Section 1. Most relevant literature studies explore single classifiers to analyse and predict patient outcome, which is not suitable for the complex nature of ICU data. From a review in Section 2, there is a lack of robust and accurate solutions in this domain. Therefore, this project utilises both single base classifiers and several ensemble techniques with and without feature selection to improve on performance and achieve more accurate results.

The LGBM and stacking models yield excellent results with both models achieving a significant improvement in performance against all the other classifiers with and without feature selection. Both are sophisticated models capable of handling various data types and have an increased range of 2%-21% accuracy on other models in this study. This is in line with research from El-Rashidy et al. (2020) where the stacking model outpreformed all other classifiers with an AUC of 0.93. Similarly, the stacking model constructed by Ghorbani et al. (2020) also achieved excellent results in prediction of ICU mortality compared to other machine and deep learning models with AUC = 0.76. However, the findings from this study exceed the current research results (AUC = 0.97). Stacking is an impressive model due to it's ability to learn how to best combine the predictions from contributing ensemble models. However, the stacking model did not preform as well with feature selection indicating that it is not as reliable or robust as the LGBM.

There is little current literature related to the use of LGBM in prediction of ICU patient outcomes due to it's relatively new release, making this a novel aspect of this research. However it is a very powerful and fast model due to its novel feature Gradient-based One-Side Sampling (GOSS) which allows advanced subsampling of the data. LGBM has the advantage of speed and ease of implementation. However, there is a risk of over-fitting with LGBM as it is highly dependant on the hyperparameter tuning so this must be taken into consideration. A criticism of this study is it does not examine the use of LGBM as a base classifier for the stacking method which could yield even more impressive results. This could be addressed in future work.

The RF model produces consistently good results over both experiments which concurs with the findings from several studies including Kong et al. (2020) and Rayan et al. (2021). This confirms the reliability of RF and it's ability to preform well with noisy and unstructured data without overfitting. Similar to Kim et al. (2011), DT performed very well against the other single models which an AUC = 0.93. Both NB and LR preformed poorly in comparison to the other models, which can be expected due to the nature of the dataset and concurs with results from both Veith and Steele (2018) and Pirracchio et al. (2015). This could be due to highly correlated features in the dataset and perhaps suggests that the correlation threshold used could be set lower. The poorer results from LR indicate a lack of linearity between the response and predictor variables. Although

SVM is a very powerful algorithm, it did not achieve the desired results due to Linear SVM being implemented as a result of computational time restraints. As with LR, the Linear SVM assumes linearity in data.

Many studies reviewed had little focus on the imbalanced nature of ICU data making it difficult to predict patient mortality accurately. This project handles the significant class imbalance which allows more effective classification. The impressive overall results of this research highlights the importance of rigorous data pre-processing, in particular addressing missing and class imbalance. Every study reviewed with the appropriate pre-processing steps implemented achieved significantly higher results than those without, as detailed by several experiments with and without pre-processing by Awad et al. (2017). The experiments that handled class imbalance and missing values achieved better results than those without. Similarly, Ghose et al. (2015) achieved an RF accuracy of 0.87 in a similar dataset used in this study without any data pre-processing in comparison to RF accuracy of 0.95 achieved in this project. The use of optimal hyperparmeters also contributed to the excellent results in this study which coincides with the results of El-Rashidy et al. (2020).

Overall, it was found that the use of feature selection did not improve classifier performance and produced consistently lower results than the full features dataset, which correspondences with ensemble methods with feature selection results found by El-Rashidy et al. (2020). Although feature selection commonly leads to improved results, there could be several reasons for the poorer results when applied to this dataset. Predictors in clinical datasets are often highly correlated with one another as certain tests depend on the patients having certain medical conditions and therefore not all tests are preformed on every patient. This means some features depend on other features due to differences in patient condition and continuous patient monitoring. Therefore, automatic feature selection may not be suitable for this study and initial feature selection by domain knowledge as examined by El-Rashidy et al. (2020) and Awad et al. (2017) may be the preferred method. It was found by Chu et al. (2012) that data-drive feature selection without prior domain knowledge did not improve results because there was no additional information in the training samples from which those feature selection methods can extract as the features were highly correlated. However, this expert medical knowledge was not available in this research which is a limitation to this study.

Another reason for the results could be that the automated feature selection method chosen may not have been as powerful as alternatives such as GA or wrapper methods on this large clinical dataset. Although, feature selection decreases the size of a dataset and computational time, including only the 20 features only could be too restrictive and may have removed vital information form the dataset. Some models used in this research, such as RF, are designed to be robust to features that are not informative of the target variable and therefore feature selection is perhaps not necessary for these models. This may produce poorer results due to unnecessary steps that add noise and cause information loss. The best option may be to use regularization methods that guard against overfitting and perform feature selection intrinsically instead of applying specific feature selection methods. This could be used in conjunction with initial domain knowledge to select the most important features. These options could be researched fully in the future work.

In terms of feature importance ranking as set out in objective 4 in Section 1, the top 20 features are ranked by importance to the response variable in Figure 5. There is a mix of demographic features and vitals in the list indicating that no particular grouping of features is dominant in prediction of mortality. BMI, weight and time spend in the

hospital prior to ICU admission are significant factors in the prediction of ICU mortality which is consistent with the research of Sanaie et al. (2021) and Monteiro et al. (2020). Those with higher BMI and weight have a greater risk of death in ICU, however it is important to note these factors have a directly proportional relationship. Some of the other top features include those that may be highly correlated with the response variable such as 'apache_4a_hosiptal_death' which suggest that some of these features should have been removed from the dataset and could be skewing results. With the current Covid-19 pandemic causing a severe strain on ICU resources, having a list of high risk mortality factors is be vital in ICU resource allocation. This can facilitate a risk assessment for each patient on ICU admission, and those with the top risk indicators could have more efficient resources allocated to them. As this dataset was collected pre Covid-19, there is no feature to indicate those patients in ICU with a positive diagnosis of Covid-19, which is something that could be investigated in future work with an updated dataset.

The final model selected is the LGBM without feature selection. This model considers 85 features from vital signs, patient demographics, and lab tests and underwent extensive data pre-processing to remove missing features and class imbalance. The proposed LGBM ensemble model achieved the best performance in all evaluation metrics and outperformed the state-of-the-art models that achieved inferior results.

8 Conclusion and Future Work

This paper proposes a novel and reliable LGBM based ensemble model for prediction of ICU mortality based on patients medical history data. A comprehensive critique of the current relevant literature is examined in Section 2. The gaps in current literature have been addressed such as lack of data preprocessing and the computation of a heterogeneous and accurate ensemble classifier for ICU mortality prediction has been examined. The large raw dataset underwent rigorous pre-processing to handle imbalanced and missing data to ensure better modelling. It was found that the application of an automated feature selection method without initial domain knowledge did not improve model results and highlights the importance of medical knowledge in feature selection for clinical datasets. The top 20 most important features in predicting ICU mortality were ranked using feature selection to indicate high risk patient mortality factors. A detailed and reproducible implementation of all single and ensemble classifiers is examined in Section 5. An exhaustive evaluation of all models is included in Section 6 and an extensive discussion of all results in regards to current literature is provided in Section 7.

Although both the LGBM and the stacking model performed equally well on the test set without feature selection, the LGBM outperformed the stacking model in the test set with feature selection indicating a more robust and reliable model. To the best of the author's knowledge LGBM has not been used to predict ICU patient mortality previously, making it a novel study that outpreformed the current state of the art. By producing a model that can aid ICU medical professionals in decision making, this allows medical professionals to make more effective data driven decisions to provide improved patient resourcing and reduce overall healthcare costs. By ranking the high risk features in patient mortality, medical professionals can identify those patients most at risk and monitor them appropriately. Computation power and time were a limitation in this study, with some compromises on performance made to save computational time. The author's lack of medical knowledge regarding the features of the dataset was also a limitation of the study.

Future work could involve the use of different, more appropriate, feature selection methods such as GA or wrapper methods to improve upon the poorer feature selection results in this study. Incorporating a medical domain expert could help to create feature subsets for better results and understanding of results. Further examination of stacking with different base classifiers could be examined. Deep learning algorithms such as ANN and CNN could be utilised to handle the time series data and comparison of the models used in this study to current scoring systems such as APACHE could be investigated.

9 Acknowledgments

This project would not be possible without the help and guidance of my supervisor Dr. Majid Latifi. I would also like to thank my parents Eamonn and Elsie, my sister Karyn and all my friends and colleagues for all their support while conducting this research.

References

Ahmed, F. S., Ali, L., Joseph, B. A., Ikram, A., Ul Mustafa, R. and Bukhari, S. A. C. (2020). A statistically rigorous deep neural network approach to predict mortality in trauma patients admitted to the intensive care unit, *Journal of Trauma and Acute Care Surgery* **89**(4): 736–742.

Alves, T., Laender, A., Veloso, A. and Ziviani, N. (2019). Dynamic prediction of icu mortality risk using domain adaptation, *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018* .

Awad, A., Bader-El-Den, M., McNicholas, J. and Briggs, J. (2017). Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach, *International Journal of Medical Informatics* **108**.

Bennett, C. E., Wright, R. S., Jentzer, J., Gajic, O., Murphree, D. H., Murphy, J. G., Mankad, S. V., Wiley, B. M., Bell, M. R. and Barsness, G. W. (2019). Severity of illness assessment with application of the APACHE IV predicted mortality and outcome trends analysis in an academic cardiac intensive care unit, *Journal of Critical Care* **50**: 242–246.

Bhattacharya, S., Rajan, V. and Shrivastava, H. (2017). Icu mortality prediction: A classification algorithm for imbalanced datasets, *31st AAAI Conference on Artificial Intelligence, AAAI 2017* .

Che, Z., Purushotham, S., Khemani, R. and Liu, Y. (2016). Interpretable deep models for icu outcome prediction, *AMIA ... Annual Symposium proceedings. AMIA Symposium* **2016**.

Chu, C., Hsu, A. L., Chou, K. H., Bandettini, P. and Lin, C. P. (2012). Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images, *NeuroImage* **60**(1): 59–70.

Darabi, H. R., Tsinis, D., Zecchini, K., Whitcomb, W. F. and Liss, A. (2018). Forecasting mortality risk for patients admitted to intensive care units using machine learning, **140**.

El-Rashidy, N., El-Sappagh, S., Abuhmed, T., Abdelrazek, S. and El-Bakry, H. M. (2020). Intensive care unit mortality prediction: An improved patient-specific stacking ensemble model, *IEEE Access* **8**.

Ferreira, F. L., Bota, D. P., Bross, A., Mélot, C. and Vincent, J. L. (2001). Serial evaluation of the sofa score to predict outcome in critically ill patients, *Journal of the American Medical Association* **286**.

Ghorbani, R., Ghousi, R., Makui, A. and Atashi, A. (2020). A new hybrid predictive model to predict the early mortality risk in intensive care units on a highly imbalanced dataset, *IEEE Access* **8**.

Ghose, S., Mitra, J., Khanna, S. and Dowling, J. (2015). An Improved Patient-Specific Mortality Risk Prediction in ICU in a Random Forest Classification Framework, **214**: 56–61.

Gilani, M., Razavi, M. and Azad, A. (2014). A comparison of simplified acute physiology score ii, acute physiology and chronic health evaluation ii and acute physiology and chronic health evaluation iii scoring system in predicting mortality and length of stay at surgical intensive care unit, *Nigerian Medical Journal* **55**.

Guo, C., Liu, M. and Lu, M. (2021). A Dynamic Ensemble Learning Algorithm based on K-means for ICU mortality prediction, *Applied Soft Computing* **103**: 107166.

Hou, W., Zhao, Z., Chen, A., Li, H. and Duong, T. Q. (2021). Machining learning predicts the need for escalated care and mortality in covid-19 patients from clinical variables, *International Journal of Medical Sciences* **18**.

Hsieh, Y. Z., Su, M. C., Wang, C. H. and Wang, P. C. (2014). Prediction of survival of icu patients using computational intelligence, *Computers in Biology and Medicine* **47**.

Kaier, K., Heister, T., Wolff, J. and Wolkewitz, M. (2020). Mechanical ventilation and the daily cost of icu care, *BMC Health Services Research* **20**.

Kim, S., Kim, W. and Woong Park, R. (2011). A comparison of intensive care unit mortality prediction models through the use of data mining techniques, *Healthcare Informatics Research* **17**(4).

Kim, S. Y., Kim, S., Cho, J., Kim, Y. S., Sol, I. S., Sung, Y., Cho, I., Park, M., Jang, H., Kim, Y. H., Kim, K. W. and Sohn, M. H. (2019). A deep learning model for real-time mortality prediction in critically ill children, *Critical Care* **23**.

Kong, G., Lin, K. and Hu, Y. (2020). Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU, *BMC Medical Informatics and Decision Making* **20**(1): 251.

Lee, J., Dubin, J. A. and Maslove, D. M. (2016). *Secondary Analysis of Electronic Health Records*, 1st edn, Springer.

Monteiro, A. C., Suri, R., Emeruwa, I. O., Stretch, R. J., Cortes-Lopez, R. Y., Sherman, A., Lindsay, C. C., Fulcher, J. A., Goodman-Meza, D., Sapru, A., Buhr, R. G., Chang,

S. Y., Wang, T. and Qadir, N. (2020). Obesity and smoking as risk factors for invasive mechanical ventilation in COVID-19: A retrospective, observational cohort study, *PLOS ONE* **15**(12): e0238552.

Nie, X., Cai, Y., Liu, J., Liu, X., Zhao, J., Yang, Z., Wen, M. and Liu, L. (2021). Mortality Prediction in Cerebral Hemorrhage Patients Using Machine Learning Algorithms in Intensive Care Units, *Frontiers in Neurology* **11**: 610531.

Pirracchio, R., Petersen, M. L., Carone, M., Rigon, M. R., Chevret, S. and van der LAAN, M. J. (2015). Mortality prediction in the icu: can we do better? results from the super icu learner algorithm (sicula) project, a population- based study, *Lancet Respir Med* **3**.

Raghupathi, W. and Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential, *Health Information Science and Systems* **2**.

Rayan, Z., Alfonse, M. and Salem, A.-B. M. (2021). An Ensemble Model for Early Sepsis Prediction using Clinical Records from the Intensive Care Unit (ICU), *Egyptian Computer Science Journal* **45**(1).

Ryan, L., Lam, C., Mataraso, S., Allen, A., Green-Saxena, A., Pellegrini, E., Hoffman, J., Barton, C., McCoy, A. and Das, R. (2020). Mortality prediction model for the triage of covid-19, pneumonia, and mechanically ventilated icu patients: A retrospective study, *Annals of Medicine and Surgery* **59**.

Sanaie, S., Hosseini, M.-S., Karrubi, F., Iranpour, A. and Mahmoodpoor, A. (2021). Impact of Body Mass Index on the Mortality of Critically Ill Patients Admitted to the Intensive Care Unit: An Observational Study, *Anesthesiology and Pain Medicine* **11**(1): 1–6.

Silva, A. D., Correia, C. C., Salgado, C. M., Finkelstein, S., Celi, L. M., Sousa, J. M. and Vieira, S. M. (2018). Fuzzy modeling for predicting patient survival rate in icu with aki, *IEEE International Conference on Fuzzy Systems* **2018-July**.

van Doorn, W. P., Stassen, P. M., Borggreve, H. F., Schalkwijk, M. J., Stoffers, J., Bekers, O. and Meex, S. J. (2021). A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis, *PLoS ONE* **16**.

Veith, N. and Steele, R. (2018). Machine learning-based prediction of icu patient mortality at time of admission, *ACM International Conference Proceeding Series* .

Xu, J., Zhang, Y., Zhang, P., Mahmood, A., Li, Y. and Khatoon, S. (2017). Data mining on icu mortality prediction using early temporal data: A survey, *International Journal of Information Technology and Decision Making* **16**.

Zheng, H. and Shi, D. (2018). Using a lstm-rnn based deep learning framework for icu mortality prediction, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11242 LNCS**.

Zimmerman, J. E., Kramer, A. A., McNair, D. S. and Malila, F. M. (2006). Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients, *Critical Care Medicine* **34**(5).