Early Hospital Mortality Prediction of Intensive Care Unit Patients Using an Ensemble Learning Approach

Aya Awad, Mohamed Bader-El-Den, James McNicholas and Jim Briggs

University of Portsmouth, Buckingham Building, Lion Terrace, Portsmouth PO1 3HE, UK Email: mohamed.bader@port.ac.uk

James McNicholas

Critical Care Unit, Queen Alexandra Hospital Portsmouth Hospitals, NHS Trust UK.

Abstract

Background: Mortality prediction of hospitalized patients is an important problem. Over the past few decades, several severity scoring systems and machine learning mortality prediction models have been developed for predicting hospital mortality. By contrast, early mortality prediction for intensive care unit patients remains an open challenge. Most research has focused on severity of illness scoring systems or data mining (DM) models designed for risk estimation at least 24 or 48 hours after ICU admission.

Objectives: This study highlights the main data challenges in early mortality prediction in ICU patients and introduces a new machine learning based framework for Early Mortality Prediction for Intensive Care Unit patients (EMPICU).

Materials and methods: The proposed method is evaluated on the Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) database. Mortality prediction models are developed for patients at the age of 16 or above in Medical ICU (MICU), Surgical ICU (SICU) or Cardiac Surgery Recovery Unit (CSRU). We employ the ensemble learning Random Forest (RF), the predictive Decision Trees (DT), the probabilistic Naive Bayes (NB) and the rule-based Projective Adaptive Resonance Theory (PART) models.The primary outcome was hospital mortality. The explanatory variables included demographic, physiological, vital signs and laboratory test variables. Performance measures were calculated using cross-validated area under the receiver operating characteristic curve (AUROC) to minimize bias. 11,722 patients with single ICU stays are considered. Only patients at the age of 16 years old and above in Medical ICU (MICU), Surgical ICU (SICU) or Cardiac Surgery Recovery Unit (CSRU) are considered in this study.

Results: The proposed EMPICU framework outperformed standard scoring systems (SOFA, SAPS-I, APACHE-II, NEWS and qSOFA) in terms of AUROC and time (i.e at 6 hours compared to 48 hours or more after admission).

Discussion and conclusion: The results show that although there are many values missing in the the first few hour of ICU admission, there is enough signal to effectively predict mortality during the first 6 hours of admission. The proposed framework, in particular the one that uses the ensemble learning approach - EMPICU Random Forest (EMPICU-RF)offers a base to construct an effective and novel mortality prediction model in the early hours of an ICU patient admission, with an improved performance profile.

Keywords: Intensive care, mortality prediction, Classification, Class Imbalance.

1. Introduction

Hospitals are subject to multiple pressures, including limited funds and healthcare resources. The intensive care unit (ICU) in particular has drawn considerable attention from the medical community due to its critically ill patients and costly resources. The ICU patient is highly monitored using electronic equipment to

measure physiological data, which provides a rich opportunity for valuable clinical data analysis. Mortality prediction for ICU patients is critical by nature as the quicker and more accurate the decisions taken by intensivists, the more the benefit for both patients and health care resources.

Two established models attempt early mortality prediction for ICU patients; they are the Mortality Probability Model (MPM) [1] and SAPS III[2]. These will be discussed thoroughly in Section II, however, they are purely statistically derived, unlike our model. Other models are designed to predict mortality after the first 24, 48 or 72 hours of ICU admission [3, 4, 5, 6, 7, 8, 9]. Even models, such as the one proposed by Calvert et al. [10] attempts to predict mortality 12 hours before in-hospital death; this study shows strong predictive accuracy but we question the practical utility of the tool, which predicts at a point twelve hours from the sampling. It is not clear at what stage in the evolution of a critical care episode that this tool should be employed to best effect. If it were used continuously until such time as a death were felt to be very high risk, there would, for many, already have been a protracted ICU course with the attendant burdens of treatment. Whilst this delay is acceptable where the intended purpose is unit quality benchmarking, it is slow for the purpose of decision assist. In contrast, the model proposed in this study attempts to predict in-hospital mortality shortly after ICU admission. It is our hypothesis that accurate prediction of hospital mortality is possible using data collected in the earliest phase of admission. Early mortality prediction is motivated by the intention to assist clinicians and patients in the assessment of the risks and benefits attending intensive care admission. We hold that it is in the interests of patients, or their advocates, to be informed of a quantitative mortality risk, as early as possible, and preferably before committing to burdensome critical care interventions, whenever that is possible.

Many studies show that customized models perform better than traditional scores. Lee et al. [11] conducted a retrospective analysis using data from the MIMIC-II database; the study concluded that custom models based on ICU-specific data provided better mortality prediction than traditional SAPS scoring using the same predictor variables. However, ICU is a very complex environment where patients may suffer from more than one condition, which makes it difficult to specify which customized model to use. Therefore, there is a need for general mortality prediction models, which is the focus of this study.

This study aims to investigate the use of DM classification methods in developing an early mortality prediction model to assist clinicians in decision making. We do this by analysing different medical variables for patients from the first six hours after ICU admission, rather than the typical 24, 48 or 72. We hypothesize that an early mortality prediction model, could help provide intensivists with a systematic interpretation of a patients observations sooner than with current methods. We define 'early' as at approximately six hours after ICU admission. The explanatory variables include demographic, vital signs and laboratory test variables. The primary outcome is hospital mortality, which is defined as death inside the hospital; we seek to identify those patients at high risk of dying inside the hospital.

Given the ICU patients early medical record (data available from the first few hours after admission), can data mining methods help in predicting the patients who are most likely to die inside the hospital? What are the most important medical attributes to consider in early mortality prediction? What are the most effective data mining methods for early prediction of hospital mortality for ICU patients?

The objectives of this study are:

1. Identify the main data mining challenges in early hospital mortality prediction for ICU patients.
2. Design a general framework for early mortality prediction for ICU patients to tackle the data mining challenges identified in point 1.
3. Evaluate and compare the performance of different attribute selections and data mining methods on a freely available ICU database.

This paper is organized as follows: section II introduces previous work that has been done in ICU patients' mortality prediction, section III introduces the general framework of early mortality prediction for ICU patients (EMPICU) presented in this research, section IV displays the dataset used and the pre-processing conducted, section V presents experiments' methods and results, section VI discusses the results and finally section VII concludes the work done in this research.

2. Related Work in Mortality Prediction for ICU patients

This section provides an overview on the ICU environment. Similar solutions for mortality prediction, including severity scoring systems and data mining approaches are reviewed. In addition some data mining challenges in mortality prediction facing medical doctors and data scientists are introduced.

2.1. Overview of the ICU environment

The intensive care unit is a complex and information rich department. Patients admitted to ICUs require close and continuous monitoring due to high illness severity and the potential for rapid disease progression. ICU patients are also heterogeneous, often suffering multiple concurrent problems, and fewer in number than patients presenting to single system specialties. Research is therefore necessarily limited by both heterogeneity and small sample numbers. For this and other reasons, the evidence base for critical care practice is less well developed than for some other acute specialties. The unique combination of rich data sources from monitoring, and a complex, heterogeneous patient population, makes the ICU setting particularly well suited for the implementation of an assistant data-driven system which analyzes large amounts of raw data that could be overlooked by human experts [12]. The use of ICU data in early prediction of mortality is an attractive open area for investigation, both for reasons of quality and cost.

2.2. Scoring systems for mortality prediction

2.2.1. Traditional scoring systems for mortality prediction

In this section, we will discuss the following traditional ICU scoring systems: (1) Acute Physiology and Chronic Health Evaluation (APACHE) [13], (2) Simplified Acute Physiology Score (SAPS) [14], (3) Sequential Organ Failure Assessment (SOFA) [15], (4) quick Sepsis-related Organ Failure Assessment (qSOFA) score [16] and (5) National Early Warning Score (NEWS) [17].

Several publications in the literature have discussed and compared mortality prediction models for ICU patients that rely on panels of experts or statistical models, also referred to as regression models[18, 13, 14, 19, 1, 20, 21, 22]. For example, APACHE [13] and SAPS [14] assess disease severity to predict outcome. The objective of these models is to characterize disease severity from patient demographics and physiological variables obtained within the first 24 hours after ICU admission in order to assess ICU performance. The models have been refined for use within specified geographical areas, such as France, Southern Europe and Mediterranean countries, and to Central and Western Europe [23, 24, 25, 2, 5, 21, 22]. Using a very different strategy, Hoogendoorn et al. [26] built two prediction models. The methods used were: (1) extraction of high-level (temporal) features from Electronic Medical Records (EMRs) and to build a predictive model; (2) definition of a patient similarity metric with prediction based on the outcome observed for similar patients. Neither approach gave optimal discrimination but the first model, using temporal features (AUROC 0.84), was superior to the patient similarity model (AUROC 0.68).

Prediction systems have evolved since their inception, but have not always led to improved discrimination. APACHE III [23] was developed in 1991 and in 2002/2003 APACHE IV [27] was developed, which provides length of stay prediction equations, in addition to the prediction capability of earlier iterations. A more detailed comparison of the current APACHE scoring systems is available in [21]. Research in [24] introduced an expanded SAPS II by adding six admission variables: age, gender, length of pre-ICU hospital stay, patient location before ICU, clinical category and presence of drug overdose. Results show that the expanded SAPS II performed better than the original and a customized SAPS II, with an AUROC of 0.879. However, a study conducted by Gilani et al. [22] comparing APACHE scores and SAPS II score, showed that the discrimination of APACHE II (as measured by the AUROC) was excellent (AUROC: 0.828) and acceptable for APACHE III (AUROC: 0.782) and SAPS II (AUROC: 0.778) scores. In addition [28] found that the discrimination of APACHE IVa was superior with (AUROC: 0.88) compared with Mortality Probability Model (MPM) III [29] (0.81) and ICU Outcomes Model/National Quality Forum (0.80) [28].

Another traditional scoring systems is the SOFA score [15], which is limited to 6 organ systems by looking at respiration, coagulation, liver, cardiovascular, central nervous system, and renal measurements. For each organ system, the score provides an assessment of derangement between 0 (normal) and 4 (highly deranged).

In addition, the qSOFA score [16] is a bedside tool that was recommended for use by the recent Third International Consensus Definitions Task Force [30] to identify high-risk patients outside the ICU. qSOFA was found to be more accurate than the systemic inflammatory system criteria (SIRS) for predicting mortality and intensive care unit (ICU) transfer in patients outside the ICU. However, the qSOFA score has yet to be validated outside of the original publication and has not been compared to early warning scores already in widespread use. The qSOFA criteria were defined as systolic blood pressure greater than 100mm Hg, respiratory rate greater than 22 breaths per minute, and altered mental status (defined as either a Glasgow Coma Scale score less than13 or an Alert Voice Pain Unresponsive scale (AVPU) other than Alert) [31].

The Royal College of Physicians recommends the use of NEWS for the routine clinical assessment of all adult patients. NEWS is calculated based on previously published tables in [17]. It is important to emphasize that in addition to being an early warning score to escalate care, NEWS has the capability of predicting mortality. NEWS has proven to perform better than 33 other systems to predict mortality within 24 hours of hospital admission [17]. We tend to compare performance of mortality prediction of ICU patients using NEWS, APACHE, qSOFA, SOFA, and SAPS at different time intervals after ICU admission. A comprehensive survey on mortality prediction in ICU can be found in [32].

According to the clinical review conducted by Vincent et al. [21], the different types of score should be seen as complementary, rather than competitive and mutually exclusive. Scoring systems have focused on providing increasingly refined methods for benchmarking ICU performance, and have laid the foundation for robust systems of quality control, but the use of such tools for individual decision assist, remains unproven.

2.2.2. Early scoring systems for mortality prediction

The Mortality Probability Model (MPM) was described by Lemeshow et al. in 1985 [33]. At admission, 137 variables were collected and 75 at 24 hours after admission. Using statistical techniques the relative importance of each variable was determined and only those with a strong association with outcome retained. This resulted in 7 variables collected at admission and 7 at 24 hours. Unlike APACHE and SAPS, this model could be applied at the time of admission. Further, the physiological variables are recorded as affirmative or negative rather than as an actual number. Lemeshow published an updated form of the model, the MPM II in 1993 [1]. This resulted in two models, $mpm0$ at admission $mpm24$ at 24 hours. $mpm0$ requires the collection of 15 and $mpm24$ a further 8 variables. Both models were shown to be good systems for reliably estimating hospital mortality. At that time $mpm0$ was, by definition, the only model for estimating hospital mortality which was independent of treatment.

The objective of the development of SAPS-III [2] was the evaluation of the effectiveness of ICU practices; therefore the focus of the model was on data available at ICU admission or within a day of admission. Missing values were coded as the reference or normal category for each variable. When data collection was used maximum and minimum values recorded during a certain time period, missing maximum values of a variable were replaced by the minimum and vice versa. Some regression imputations were performed if noticeable correlations to available values could be exploited. Selection of variables was done according to their association with hospital mortality, together with expert knowledge and definitions used in other severity of illness scoring systems. The objective of using this combination of techniques rather than regression-based criteria alone was to reach a compromise between over-sophistication of the model and knowledge from sources beyond the sample with its specific case mix and ICU characteristics. In the study conducted in [19], the authors compared the predictive ability of the Simplified Acute Physiology Score (SAPS) II and SAPS 3 (originally developed from data collected in 19911992 and 2002, respectively) on a sample of critically ill patients. Both scores provided unreliable predictions, but unexpectedly the newer SAPS 3 turned out to overpredict mortality more than the older SAPS II.

Among the previously discussed traditional scoring systems only few models are designed for early mortality are suitable for early mortality prediction in ICU (e.g. Mortality Probability Model). However, these models are not widely used due to their low discrimination power. Moreover, many of required attributes are not always available at ICU admission. As a result, this triggered the need for a different approach for earlier mortality prediction.

2.3. Data mining techniques for mortality prediction

Various authors have advocated the use of machine learning techniques over the use of logistic regression methods for predicting ICU mortality. Research in [34] and [35] has reported better performance by Artificial Neural Networks (ANN) over logistic regression. However, research in [36, 37, 38] found that logistic regression and neural networks performed similarly for mortality prediction. Others [6, 7, 39, 8, 9] found that Decision Trees (DTs) and Support Vector Machines (SVMs) performed better than ANN and logistic regression.

In 2011, Ribas et al. [6] showed that the use of SVMs resulted in better prediction accuracy compared with the APACHE II score. Likewise, a study conducted by Kim et al. [7] compared the predictive accuracy of ANN, SVM and DT derived from the University of Kentucky Hospital's ICU patients' data with the APACHE III scoring system. Results showed that the best performing model is Clementine's C5.0 algorithm (DT) followed by SVM, APACHE III and ANN. These results confirm earlier findings in Delen et al.[8] who also reported that C5.0 was the best predictor with an accuracy of 0.936 in predicting breast cancer survivability. In addition, Crawford et al. [9] concluded that a decision tree used in their study provided a clinically acceptable mining result in predicting susceptibility of prostate carcinoma patients at low risk for lymph node spread.

On the other hand, Ramon et al. [40] reported that the AUROCs of decision tree based algorithms (decision tree learning, 0.65; first order random forests, 0.81) yielded smaller areas compared to those of naive Bayesian networks (AUROC, 0.85) and tree-augmented naive Bayesian networks (AUROC, 0.82) in their study on a small dataset containing 1,548 mechanically ventilated ICU patients. Similarly Pirracchio et al. [5] reported that a Bayesian Additive Regression Tree (BART) is the best candidate when using transformed variables, while Random Forests (RF) outperformed all other candidates when using untransformed variables.

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. However, some models reveal strengths over others in certain aspects. For example, the major advantage for the use of a DT over other models lies in its descriptive modelling as it explains hidden clinical implications unlike an ANN which lacks logic between input and output nodes. From another perspective, DT, RF, ANN, Bayesian networks and SVM can handle large size data samples and integrate background knowledge into analysis [41].

3. A framework for Early Mortality Prediction for ICU patients

In this section, we present the general framework for dealing with early mortality prediction in this study as shown in figure 1. There are a number of challenges due to the characteristics of typically available ICU data. The framework address three of these: (1) missing values in data; (2) attribute selection; and (3) the class imbalance problem.

3.0.1. Attribute Selection

It is often difficult to decide which attributes in a dataset should be used to construct the model. Choosing all attributes may result in a model that is inefficient to compute or is over-fitted to the data. One data-driven approach is to select those attributes with high availability/coverage, meaning that the attribute/test should be measured at least once for each patient. Another is to base the selection on those that contribute the highest information gain is predicting outcome. A problem-driven approach is to use the expertise of ICU consultants or to select the same attributes used in related work.

3.0.2. Missing values

Not all medical variables/tests are measured for all patients within the first few hours of admission, therefore (for each patient) there may be some expected data missing. There are three types of missing values in data: (1) missing completely at random (MCAR); (2) missing at random (MAR); and (3) missing not at random (MNAR) [42]. Missing values in the ICU could be interpreted as a normal value (MNAR);

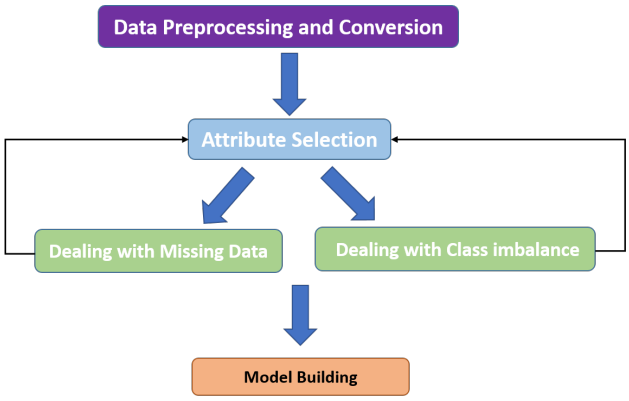

Figure 1: Proposed framework of an early mortality prediction model in the ICU

if an individual patients record has multiple entries missing, it may be explained that this is because they were regarded as being less sick than others, so they were not prioritized. Equally, the patient may have been regarded as being extremely sick, so they died before much can be done. Distinguishing these cases is not simple in the absence of other information.

Missing values can be handled either by ignoring those records from the dataset that are not complete, or by filling in missing values by one of a number of techniques. One technique is to substitute the missing value by the mean or mode value of each attribute. The Weka data mining software [43] is a collection of machine learning algorithms for data mining tasks. It also contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization. In Weka there is a filter called "ReplaceMissingValues" that permit the replacement of all missing values in a dataset using the mean of each attribute.Alternatively, the missing value can be predicted by using a learning algorithm, such as Multiple Imputation or EMImputation, which replaces missing numeric values using Expectation Maximization [44]. It is an iterative method for finding maximum likelihood estimates of parameters in statistical models.

3.0.3. Class imbalance

Class imbalance is a major problem in EMPICU, because the number of patients who die inside the hospital is relatively tiny in comparison with the number who survive. Techniques for dealing with class imbalanced datasets include modifying the dataset (re-sampling) [42, 45], making the classifier 'cost sensitive' [46] or a hybrid method that combines both. Re-sampling involves modifying an imbalanced dataset to change the imbalance ratio (majority class / minority class). There are two types of re-sampling: undersampling and oversampling. Oversampling is the technique of increasing the number of records in the minority class, while undersampling is the technique of decreasing the number of records in the majority class. In addition, classifiers can be made cost sensitive with the use of a cost matrix [46].

4. Data and Preprocessing

This section provides an overview on the MIMIC II dataset that is used for analysis and modelling in this research. Preliminary analysis and attribute selection methods are also described.

4.1. Data Description

In this study, we used the MIMIC II [47] database for analysis and modelling. In preparing the data for use, an extensive examination of data variables was conducted, which meant making a variety of choices and assumptions. In this analysis we included only patients at the age of 16 or above with a single ICU stay in either the Medical ICU (MICU), Surgical ICU (SICU) or Cardiac Surgery Recovery Unit (CSRU). This cohort included 11,722 patients. We define patient mortality as death inside the hospital.

There are two basic types of data in the MIMIC II database:

1. clinical data stored in a relational database that can be queried using Structured Query Language (SQL); and
2. bedside monitor waveforms stored in flat binary files.

There are over 25,000 patients in the relational database. Only a small fraction (around 2,500) of these patients have associated waveform records. As a consequence, the analysis in this study focuses on patient records in the relational database.

Each patient has a unique subject identifier, however each patient may have one or more associated ICU stays. There are 38,320 ICU stays in the database. Out of these stays, 33,891 ended with discharge alive from the hospital, while 4,430 ended with death inside the hospital. There are 25,665 patients with single ICU stays in the database.

A large set of candidate variables was considered from nurse-charted observations/ vital signs stored in the chartevents table, such as temperature, heart rate and blood pressure. Other candidate variables included lab tests stored in the labevents table, such as hematocrit, white blood cells count and creatinine. Categorical observations, such as ICU service type (surgical versus medical admission), disease variables (acquired immunodeficiency syndrome (AIDs) and metastatic cancer) and patient demographics, such as age and gender were also considered in the research.

4.2. Preliminary data analysis

This research focuses on both the chart and lab-test variables collected in the first 6 hours of a patients admission for the prediction of mortality. As a first step patient chart records and lab-test records within the first 6 hours were extracted. There are 38,207 patient ICU stays associated with chart attributes within the first 6 hours compared to 38,319 patient ICU stays in the dataset. Also, there are 31,175 patient ICU stays associated with lab test data within the first 6 hours compared to 38,319 patient ICU stays in the dataset. This is because very few patients didnt have chart attributes and/or lab tests recorded for them within the first 6 hours. For simplicity reasons, we combined chart data with lab tests data in one table and considered only patients with single ICU stays. This aggregation resulted in 25,665 single patient ICU stays. After filtering those patients below the age of 16 and considering only those admitted in Medical ICU (MICU), Surgical ICU (SICU) or Cardiac Surgery Recovery Unit (CSRU), this resulted in 11,722 patients with single ICU stays.

The dataset contains 4,832 chart variables and 713 lab-tests, however some of these variables/ tests are measured for only a few patients within the first 6 hours. We therefore calculated both the coverage of each chart attribute and lab-test for patients within the first 6 hours to select those variables/ tests with high coverage. We only ignored attributes with coverage below 10%. This explains why some common variables in the literature might not be included in this study as they had low coverage in the first 6 hours of admission.

In addition to the initial statistical experiments on the chart attributes and lab tests, both expertise from ICU consultants and data proposed in previous work, together with data mining algorithms were also considered in attribute selection. The following section discusses thoroughly which attributes are considered in this study.

4.3. Selected Attributes

We selected 33 chart attributes and 25 lab-tests from the initially identified attributes with high coverage. Attributes with higher coverage were considered, resulting in a total of 20 unique variables; 29 if we count maximums and minimums. The first column in Table I shows the complete list of attributes that are used in the experiments grouped by their medical category.

4.4. Confidentiality and ethical consideration

A Human Subjects Protections course called Protecting Human Research Participants was completed and its certificate (certificate number: 1765456) was earned on 18th May, 2015 as part of the MIMIC II Clinical Database Restricted Data Use Agreement. In addition, the Ethics Review certificate (certificate code: 4111-3BE3-FA5D-B16B-9059-D4FB-0D33-184C) required by Portsmouth University has been completed as well.

Table 1: shows variable names grouped by category and number of patient records used in each experiment. X means that this variable is present in the experiment, T1 refers to the corresponding attribute as being Top 1, T2 refers to the attribute coming second, T3 refers to Top 3...etc. in the ranking of attributes regarding information gain.

Attribute name	Max/Min	All	VS	Top 5	Top 10	Top 15	Top 5 (F)	Top 10 (F)	Top 15 (F)
Demographic Variable(s)									
Age		X	X	T1*	T1	T1	T1	T1	T1
Main Vital Sign(s)									
Heart Rate	Max.	X	X	T4	T4	T4	T4	T4	T4
Heart Rate	Min.	X	X	T5	T5	T5	T5	T5	T5
Systolic Blood Pressure	Max.	X	X	X	X	X			
Systolic Blood Pressure	Min.	X	X	X	T10	T10		T10	T10
Temperature (C)	Max.	X	X	X	X	T13			T13
Respiratory Rate	Max.	X	X	T3	T3	T3	T3	T3	T3
Examination Variable(s)									
Glasgow Coma Scale (GCS)	Min.	X	X	X	T9	T9		T9	T9
Lab tests Variable(s)									
Arterial Blood Oxygen	Min.	X	X	X	X	X			
Fractional Inspired Oxygen	Max.	X		X	X	X			
Serum Urea Nitrogen Level	Max.	X		T2	T2	T2	T2	T2	T2
Serum Creatinine	Max.	X	X	X	T6	T6		T6	T6
INR	Max.	X		X	X	T12			T12
INR	Min.	X		X	X	X			
Sodium Level	Max.	X		X	X	X			
Sodium Level	Min.	X		X	X	T15			T15
Potassium Level	Max.	X		X	X	X			
Potassium Level	Min.	X		X	T8	T8		T8	T8
White Blood Cells	Max.	X		X	X	X			
White Blood Cells	Min.	X		X	X	T11			T11
Bilirubin	Max.	X		X	X	X			
Bilirubin	Min.	X		X	X	X			
Platelets Count	Max.	X		X	X	X			
Platelets Count	Min.	X		X	X	T14			T14
Hematocrit	Max.	X		X	X	X			
Hematocrit	Min.	X		X	X	X			
Type of Admission/unit		X		X	T7	T7		T7	T7
Disease Variable(s)									
AIDs		X		X	X	X			
Metastatic Cancer		X		X	X	X			
Number of patient records		11,722	11,722	6,701	3,418	1,356	6,701	3,418	1,356
Age (mean)		64.339	64.339	63.984	64.32	61.426	63.984	64.32	61.42
Age (st. deviation)		22.587	22.587	21.559	17.775	19.696	21.559	17.775	19.696
Number of in-hospital deaths		1488	1488	919	409	283	919	409	283
Number of survivals		10,234	10,234	5,782	3,009	1,073	5,782	3,009	1,073
Number of Males		6,571	6,571	3,832	2,132	747	3,832	2,132	747
Number of Females		5,122	5,122	2,864	1,283	606	2,864	1,283	606

5. Early Prediction using DM Techniques

The aim of this study is to investigate the use of data mining in predicting mortality early. This research performs experimental investigation on ICU patients data using data mining classification techniques to predict early mortality. In this study, we define early as the first six hours of admission. This interval was reached after consulting several intensivists and considering gaps in literature.

This section presents the results for the top performing DM algorithms - Random Forest, Decision Trees, Naive Bayes and PART. It is important to note here that we have also evaluated a larger set of algorithms, such as Support Vector Machines (SVM) and JRip, however they were outperformed by the reported methods. Random Forest is one of the most accurate ensemble learning algorithms available. For many datasets, it produces a highly accurate classifier. It runs efficiently on large databases and it has an effective method for estimating missing data and maintaining accuracy when a large proportion of the data is missing.Decision Trees are extremely fast at classifying unknown records. They are quite robust in the presence of noise. Noise in ICU data could mean errors in data or useless attributes in prediction, so in other words Decision trees provide a clear indication of which fields are most important for prediction.PART uses partial decision trees to generate the decision list shown in the output. Only the final decision list is used in classification. The Naive Bayes algorithm affords fast, highly scalable model building and scoring. It scales linearly with the number of predictors and rows [42].

We conducted five different experiments. In experiments A and B, we evaluated each of the four data mining algorithms on each of six versions of the dataset (second column of Table II):

1. original dataset (original),
2. dataset after modified by applying the Synthetic Minority Oversampling Technique (SMOTE) [48], an oversampling technique that involves increasing the size of the minority class with the insertion of synthetic data (original+smote),
3. dataset after replacing missing values with the mean (rep1) to handle the issue of missing values
4. dataset after replacing missing values with mean and then applying SMOTE (rep1+smote),
5. datasets after replacing missing values using the EMImputation algorithm (rep2),
6. dataset after replacing missing values using EMImputation algorithm and applying SMOTE (rep2+smote).

In experiment A, the dataset contained all 20 unique variables (29 if counting maximums and minimums) listed in third column of Table I. In experiment B, the dataset contained only the eight unique vital signs variables (10 if counting maximums and minimums) listed in the fourth column of Table I.

In experiments C, D and E, we chose to eliminate patient records from the original dataset (11,722 patients) that contain missing values in key attributes. Attributes were ranked by how they contributed to information gain (IG) (i.e. those variables that contribute to better classification). In experiment C, we eliminated patient records that were missing any of the top five attributes. In experiment D, we eliminated records missing any of the top 10 attributes. In experiment E, we eliminated all records missing any of the 15 attributes. The InfoGainAttributeEval algorithm in Weka [43] evaluates the worth of an attribute by measuring the information gain with respect to the class.

In experiments C, D and E we used four versions of the dataset (second column of Table II):

1. dataset with eliminated records and the 20 unique variables (original),
2. dataset with eliminated records and the 20 unique variables and then applying smote (original+smote),
3. dataset with eliminated records and the top filtered ranked variables only (filtered top, 5, 10 and 15), and
4. dataset with eliminated records and the top filtered ranked variables only and then applying smote (filter+smote).

The number of patient records, in-hospital deaths and survivors, male and female in each experiment are shown in Table I.

All experiments were done using Weka (version 3.7.13; University of Waikato, Hamilton, New Zealand). The results noted in tables II, III and IV are AUROC of the average of 10 runs, each run is 10-fold cross-validated. The results are presented in detail in the following subsections.

Figure 2 illustrates the general framework of experiments 1 and 2. Figure 3 illustrates the framework of experiments 3, 4 and 5. Table I shows variable names and the number of patient records used in each experiment. The list of the experiments is as follows:

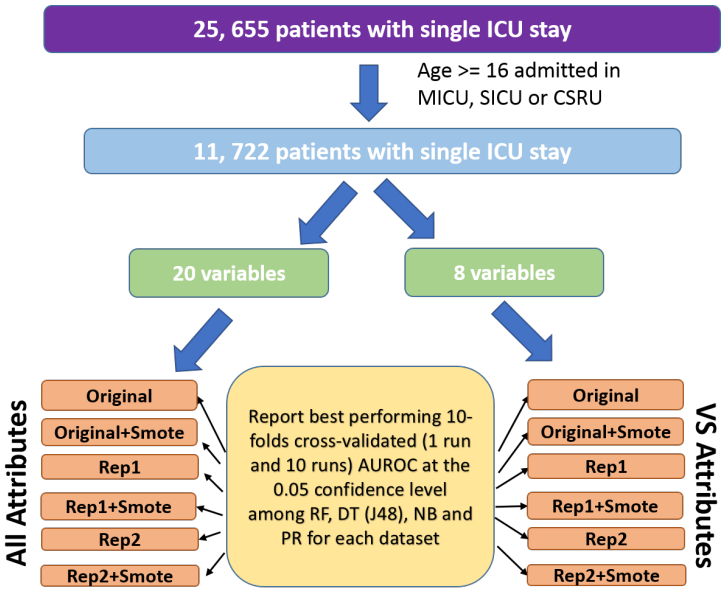

Figure 2: illustrates the general framework of experiments 1 and 2

5.1. Experiment - All 20 Selected Attributes

Methods - A total of 11,722 ICU patients and 20 variables were selected from the first six hours of a patient's admission for modelling.

Results - Table II shows the performance of the four machine-learning algorithms (at 0.05 confidence level) in predicting early hospital mortality among this patient cohort. Results were obtained on the original, original+smote, rep1, rep1+smote, rep2 and rep2+smote datasets. Among the six experiment categories, EMPICU-RF performed best, followed by EMPICU-PART, EMPICU-NB then EMPICU-DT. The most effective EMPICU-RF performance model was obtained on the rep1 and rep1+smote datasets with (AUROC = 0.85 ± 0.01).

5.2. Experiment - Vital Signs Attributes

Methods - A total of 11,722 ICU patients and 8 variables were selected from the first six hours of a patient's admission for modelling. The variables include: age in addition to 7 vital signs (temperature, heart rate, respiratory rate, systolic blood pressure, arterial blood oxygen, Glasgow Coma Scale and creatinine). Maximum temperature, maximum and minimum heart rate, maximum respiratory rate, maximum and minimum systolic blood pressure, minimum arterial blood oxygen, minimum Glasgow Coma Scale and maximum creatinine are considered.

Results - Table II shows the performance of four machine-learning algorithms (at 0.05 confidence level) in predicting early hospital mortality among this patient cohort. Results were obtained on the original, original+smote, rep1, rep1+smote, rep2 and rep2+smote datasets. Among the six experiment categories, EMPICU-RF also performed best, followed by EMPICU-PART, EMPICU-NB then EMPICU-DT. The most effective EMPICU-RF performance model was obtained on the rep1+smote dataset with (AUROC = 0.90 ± 0.01).

5.3. Experiment - Top 5 Attributes

Methods - A total of 6,701 ICU patients and 20 variables resulted from eliminating patient records in the dataset that contain missing values in any of the top 5 variables (original). The same experiment was run, but this time with filtering only the top 5 variables: age, serum urea nitrogen, respiratory rate max, heart rate max and heart rate min (filter).

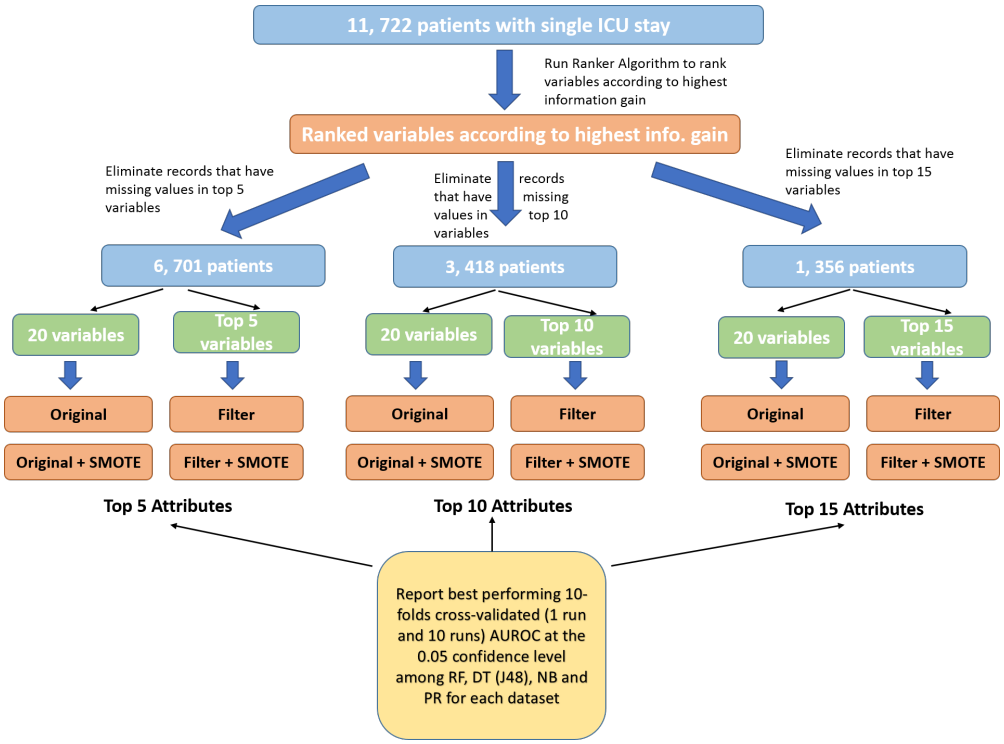

Figure 3: illustrates the general framework of experiments 3, 4 and 5

Results - Table II shows the performance of four machine-learning algorithms (at 0.05 confidence level) in predicting early hospital mortality among among this patient cohort (original) and (filter). Results were also obtained after applying SMOTE for both categories (original+smote) and (filter+smote). In the original experiment category, EMPICU-RF performed best, followed by PART, NB then DT, while in the filter experiment category NB performed best, followed by EMPICU-RF, EMPICU-PART then EMPICU-DT. The most effective EMPICU-RF performance model was obtained on the original dataset with (AUROC = 0.86 ± 0.02) and the most effective NB performance model was obtained on the filter dataset with (AUROC = 0.79 ± 0.02).

5.4. Experiment - Top 10 Attributes

Methods - A total of 3,418 ICU patients and 20 variables resulted from eliminating patient records in the dataset that contain missing values in any of the top 10 variables (original). The same experiment was run, but this time with filtering only the top 10 variables: age, serum urea nitrogen, respiratory rate max, heart rate max, heart rate min, creatinine max, care unit name, potassium min, GCS min and systolic blood pressure min (filter).

Results - Table II shows the performance of four machine-learning algorithms (at 0.05 confidence level) in predicting early hospital mortality among this patient cohort (original) and (filter). Results were also obtained after applying SMOTE for both categories (original+smote) and (filter+smote). Among the two experiment categories (original and filter), EMPICU-RF also performed best, followed by NB, PART then DT. The most effective EMPICU-RF performance model was obtained on the original and original+smote datasets with (AUROC = 0.89 ± 0.02). Figure 4 shows the performance of the 4 algorithms on the original dataset on the Yes class (patients at risk of dying inside the hospital).

5.5. Experiment - Top 15 Attributes

Methods - A total of 1,356 ICU patients and 20 variables resulted from eliminating patient records in the dataset that contain missing values in any of the top 15 variables (original). The same experiment was

Table 2: Performance of early mortality prediction models developed using 10-fold cross validated RF, DT, NB and PART in the different experiment settings measured with AUROC.

Dataset	Experment	EMPICU-RF	EMPICU-DT	EMPICU-NB	EMPICU-PART
All Attributes	Original	0.83 ± 0.02	0.69 ± 0.04	0.75 ± 0.02	0.80 ± 0.02
	Original+Smote	0.79 ± 0.03	0.73 ± 0.03	0.75 ± 0.02	0.78 ± 0.02
	Rep1	0.85 ± 0.01	0.64 ± 0.03	0.76 ± 0.02	0.76 ± 0.03
	Rep1+Smote	0.85 ± 0.01	0.65 ± 0.03	0.77 ± 0.02	0.75 ± 0.02
	Rep2	0.84 ± 0.02	0.67 ± 0.03	0.77 ± 0.02	0.78 ± 0.02
	Rep2+Smote	0.84 ± 0.02	0.67 ± 0.03	0.77 ± 0.02	0.77 ± 0.02
VS Attributes	Original	0.80 ± 0.02	0.63 ± 0.05	0.77 ± 0.02	0.78 ± 0.02
	Original+Smote	0.78 ± 0.03	0.72 ± 0.03	0.77 ± 0.02	0.77 ± 0.02
	Rep1	0.82 ± 0.02	0.70 ± 0.03	0.75 ± 0.02	0.82 ± 0.01
	Rep1+Smote	0.90 ± 0.01	0.79 ± 0.02	0.76 ± 0.02	0.84 ± 0.01
	Rep2	0.82 ± 0.02	0.71 ± 0.02	0.75 ± 0.02	0.77 ± 0.02
	Rep2+Smote	0.82 ± 0.02	0.71 ± 0.03	0.75 ± 0.02	0.77 ± 0.02
Top 5 Attributes	Original	0.86 ± 0.02	0.75 ± 0.03	0.80 ± 0.02	0.82 ± 0.02
	Original+Smote	0.85 ± 0.02	0.76 ± 0.03	0.79 ± 0.02	0.81 ± 0.02
	filter	0.78 ± 0.02	0.73 ± 0.04	0.79 ± 0.02	0.77 ± 0.02
	filter+Smote	0.78 ± 0.02	0.75 ± 0.03	0.79 ± 0.02	0.77 ± 0.03
Top 10 Attributes	Original	0.89 ± 0.02	0.71 ± 0.06	0.84 ± 0.03	0.82 ± 0.04
	Original+Smote	0.89 ± 0.02	0.73 ± 0.06	0.84 ± 0.03	0.81 ± 0.04
	filter	0.87 ± 0.03	0.71 ± 0.06	0.83 ± 0.03	0.79 ± 0.04
	filter+Smote	0.87 ± 0.03	0.72 ± 0.05	0.83 ± 0.03	0.80 ± 0.04
Top 15 Attributes	Original	0.83 ± 0.04	0.63 ± 0.07	0.78 ± 0.05	0.73 ± 0.06
	Original+Smote	0.82 ± 0.04	0.65 ± 0.06	0.77 ± 0.05	0.73 ± 0.06
	filter	0.82 ± 0.04	0.62 ± 0.07	0.78 ± 0.05	0.72 ± 0.05
	filter+Smote	0.82 ± 0.04	0.66 ± 0.06	0.78 ± 0.05	0.73 ± 0.05

run, but this time with filtering only the top 15 variables: age, serum urea nitrogen, respiratory rate max, heart rate max, heart rate min, creatinine max, care unit name, potassium min, GCS min, systolic blood pressure min, White Blood Cells min, blood clotting - INR min, Temperature max, platelets count min and sodium min (filter).

Results - Table II shows the performance of four machine-learning algorithms (at 0.05 confidence level) in predicting early hospital mortality among this patient cohort (original) and (filter). Results were also obtained after applying SMOTE for both categories (original+smote) and (filter+smote). Among the two experiment categories (original and filter), EMPICU-RF also performed best, followed by EMPICU-NB, EMPICU-PART then EMPICU-DT. The most effective EMPICU-RF performance model was obtained on the original dataset with (AUROC = 0.83 ± 0.04). Figure 5 shows the performance of the 4 algorithms on the filtered dataset on the Yes class (patients at risk of dying inside the hospital) after 6 hours of admission compared to SOFA and SAPS-I scores calculated after 24 hours of admission.

6. Comparison with traditional scoring systems

As shown in Figure 5, we compared the best performing EMPICU model, EMPICU-RF to traditional scoring systems, such as SOFA, SAPS-I, APACHE-II, NEWS and qSOFA. We used the already calculated SOFA and SAPS-I scores (after 24 hours of ICU admission) in the MIMIC-II database. We chose to calculate APACHE-II, NEWS and qSOFA scores in order to have a complete and diverse comparison considering the most effective scores we surveyed in the literature. However it was not possible to calculate MPM because some of the attributes were not available in the MIMIC-II database. As shown on the graph, the proposed model outperforms all the traditional scoring systems. Table 4 displays the AUROC and the standard deviation of each method.

Table 3: ranks top obtained results from the results shown in Table II. Results displays AUROC ± standard deviation for models developed using the most effective EMPICU model

	Experiment	AUROC
VS Attributes	Rep1+Smote	0.90 ± 0.01
Top 10 Attributes	Original	0.89 ± 0.02
Top 10 Attributes	Original+Smote	0.89 ± 0.02
Top 10 Attributes	Filter	0.87 ± 0.03
Top 10 Attributes	Filter+Smote	0.87 ± 0.03
Top 5 Attributes	Original	0.86 ± 0.02
Top 5 Attributes	Original+Smote	0.85 ± 0.02
All Attributes	Rep1+Smote	0.85 ± 0.02
All Attributes	Rep1	0.85 ± 0.01
All Attributes	Rep2	0.84 ± 0.02
All Attributes	Rep2+Smote	0.84 ± 0.02

Table 4: ranks our proposed EMPICU model and the scoring systems by best performance using AUROC

Scoring System	**AUROC**	**St. Deviation**
RF at 6 hours	0.82	0.04
SAPS at 24 hours	0.650	0.012
APACHE at 24 hours	0.650	0.017
NEWS at 24 hours	0.641	0.017
SOFA at 24 hours	0.623	0.013
qSOFA at 24 hours	0.544	0.012

7. Results' Discussion

When comparing the performance of all five experiments, we find that when using the vital signs and top 10 attributes the prediction performance is better than when using the top 5, top 15 and all 20 unique attributes. Table III ranks the experiments that showed the best performance (highest AUROC) using the best performing model, EMPICU-RF. As displayed in tables II and III, in general applying the SMOTE oversampling technique enhances the classification performance. Both replacing the missing values with mean (rep1) and replacing missing values with EMImpuation (rep2) gave almost similar performance results.

In general in the top 5, 10 and 15 experiment categories, the models developed with the original attributes (without any filtering) performed better than those with filtering. In the experiments without filtering, top 10 (original) and (original+smote) performed best (AUROC = 0.89 ± 0.02), followed by top 5 (original) (AUROC = 0.86 ± 0.02). As for the filtered experiments, top 10 (filter) and (filter+smote) also performed best (AUROC = 0.87 ± 0.03), followed by top 15 (filter) and (filter+smote) (AUROC = 0.82 ± 0.04).

When comparing the novel model with traditional scoring systems, we find that all the scoring systems are very similar in performance, which confirms what other researchers, such as [49] has mentioned regarding all traditional scoring systems developed that they handle almost similar approaches and focus on selecting the best prediction model upon which to base performance measurement. In contrast, the EMPICU model proposed rely on data mining and machine learning methods which generates more sophisticated models capable of detecting hidden patterns, deal with large amount of data and has higher discrimination power than existing traditional ICU scoirng systems.

8. Conclusion

There are several severity scoring and data mining systems that are available in the ICU. However, almost all of these models are designed to predict mortality after one or more days of admission. This paper aims to draw the attention of the medical and data science communities to the importance and the

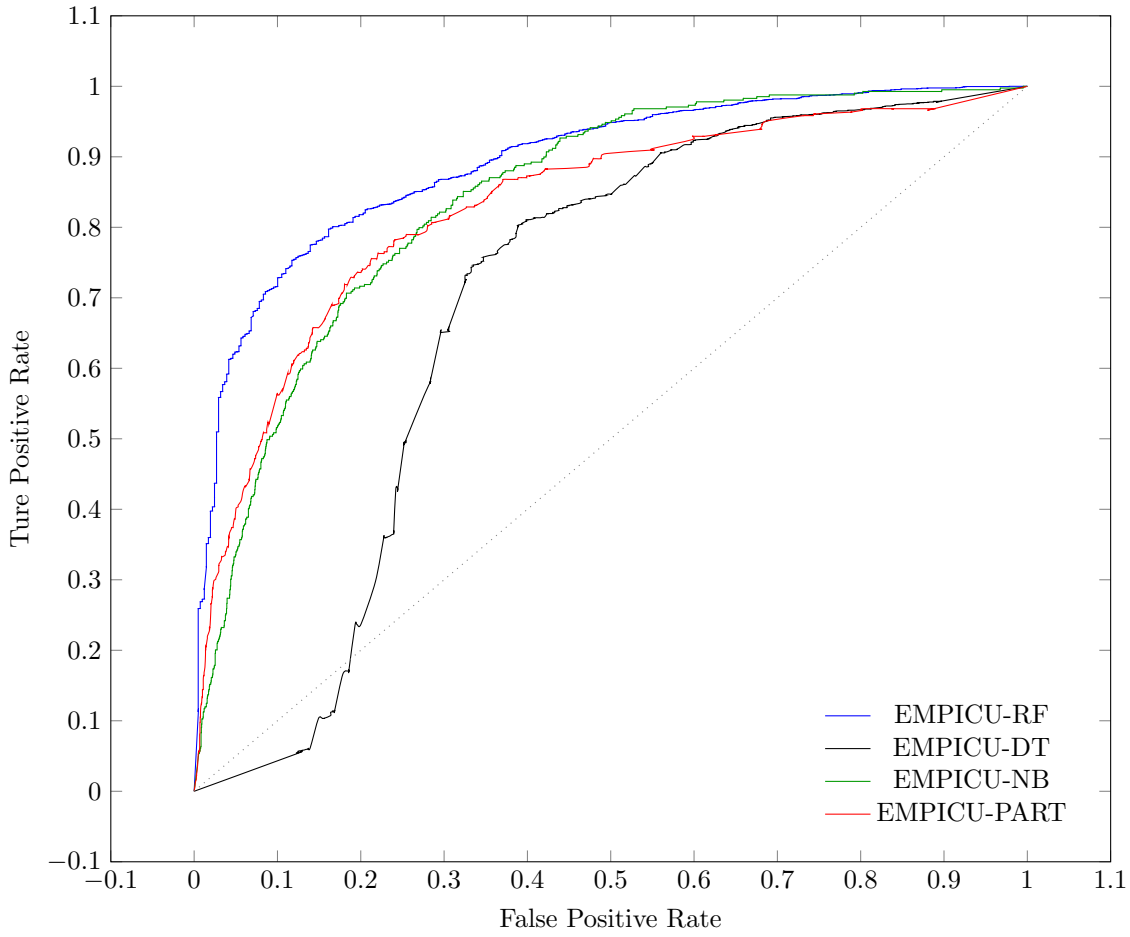

Figure 4: The performance of all EMPICU models on Top 10 Attributes dataset (Original) on the Yes class (patients at risk of dying inside the hospital)

feasibility of early mortality prediction for ICU patients. The paper presented a general framework of early mortality prediction in ICU. The framework has been evaluated on the early admissions records of 11,722 patients (from MIMIC II database) using a wide range of data mining methods. We now intend to validate this work on the MIMIC-III [50] database, which was released in August 2015, one year after our research project commenced. The contribution of the current study is both to clinical practice and for data scientists and research purposes.

From a machine-learning perspective, the most interesting outcome of this study was that the RF model outperformed those of PART, NB and DT. In the two experiments, RF performed significantly better than PART, NB and DT (at a 5% confidence level). This finding supports why in past studies, DTs are not the favoured choice of data miners. For example, Ramon et al. [40] reported that the AUROCs of a DT yielded smaller areas compared to a RF (DT, 0.65; first order RF, 0.81). Nonetheless, past studies have reported controversial finding on DTs and SVMs [8, 9, 6, 39]. Considering the limited information in literature about the use of decision-tree based algorithms for predicting health outcomes, this study contributes to our understanding of the performance of decision-tree based algorithms (RF and DT) in comparison to the NB and rule-based PART models on the MIMIC II dataset. In addition, we tried building models using the SVM and JRip algorithms, but the results were very poor and not given here.

We believe that our study has several important results:

1. Using the SMOTE oversampling technique enhanced the classification performance of all models,

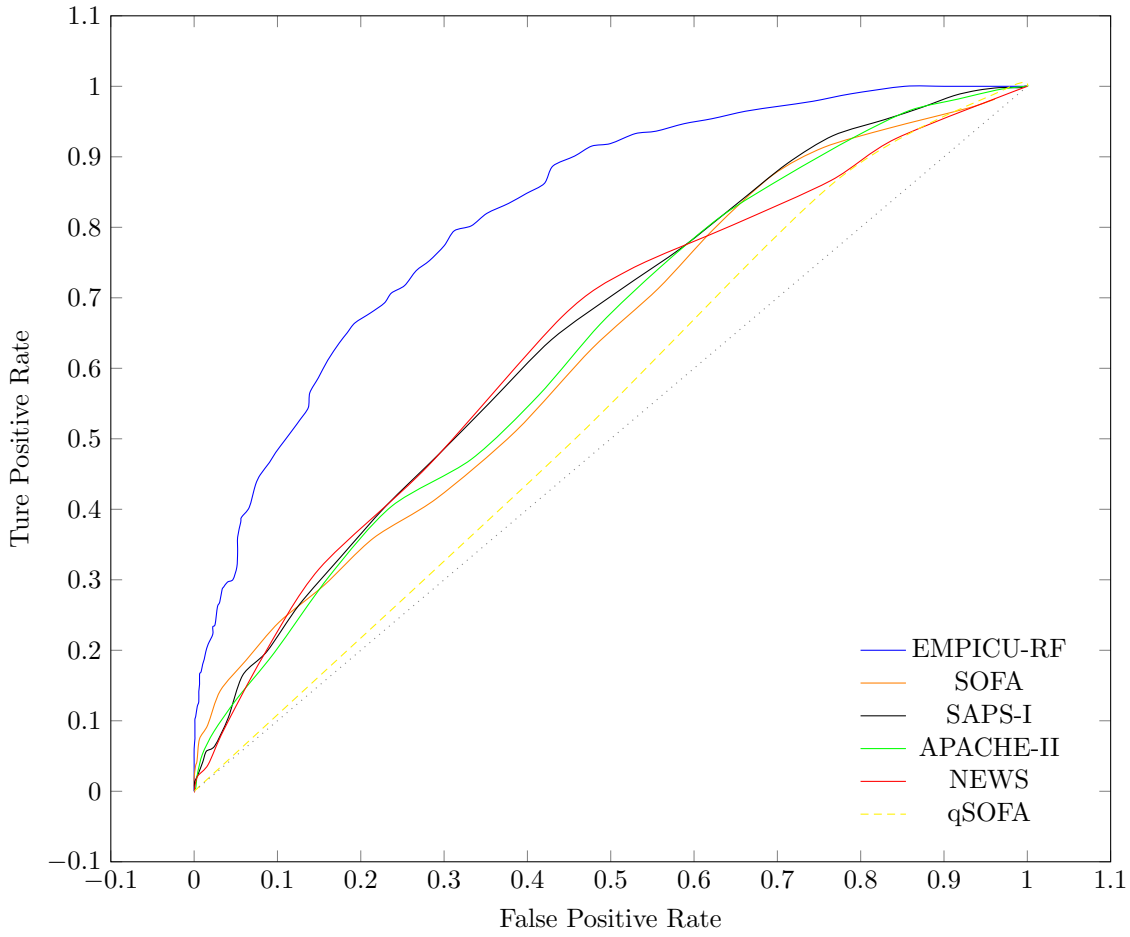

Figure 5: The performance of EMPICU-RF (superior model) on Top 15 Attributes dataset (filter) on the Yes class (patients at risk of dying inside the hospital) after 6 hours of ICU admission compared to SOFA, SAPS-I, APACHE-II, NEWS and qSOFA scores calculated after 24 hours of ICU admission

2. The best model was produced where vital signs with missing values were replaced by mean values, and SMOTE applied (rep1+smote) with AUROC = 0.90 ± 0.01,
3. Filtering out records that are missing key attributes does not contribute to better models,
4. Model performance is improved by correctly identifying the best predictive variables, not by having the largest number of variables, contrary to the findings of Pirracchio et al.; in our study the vital signs provide the best predictive model,
5. Our model compared favourably with traditional scoring systems (SOFA, SAPS-I, APACHE-II, NEWS and qSOFA) in terms of AUROC and time (i.e at 6 hours compared to 24, 48 or 72 after admission)

For clinicians, we believe that the major value of this research is to demonstrate an effective methodology, from which a clinically valuable prediction tool may subsequently be developed. We have not considered the impact of a predictive tool on clinical decision making, nor how such an impact might be measured, and we acknowledge that these are important matters, but we have shown that modern data science may be used effectively with an intensive care database, for early prediction of a clinically meaningful outcome. It is our intention that this work will be refined, to improve predictive accuracy, in the hope that it will eventually become of direct value both to clinicians and patients, to empower decision making early in critical care admission. We envisage the development of a generic system, which would be suitable for local institutional adoption and carefully controlled integration into local systems, alongside conventional ICU performance

benchmarking and whole hospital mortality monitoring.

9. Declaration of interest

Conflict of interest: none.

10. References

[1] S. Lemeshow, D. Teres, J. Klar, J. S. Avrunin, S. H. Gehlbach, J. Rapoport, Mortality probability models (mpm ii) based on an international cohort of intensive care unit patients, Jama 270 (20) (1993) 2478–2486.

[2] R. P. Moreno, P. G. Metnitz, E. Almeida, B. Jordan, P. Bauer, R. A. Campos, G. Iapichino, D. Edbrooke, M. Capuzzo, J.-R. Le Gall, et al., Saps 3from evaluation of the patient to evaluation of the intensive care unit. part 2: Development of a prognostic model for hospital mortality at icu admission, Intensive care medicine 31 (10) (2005) 1345–1355.

[3] Y. Luo, Y. Xin, R. Joshi, L. Celi, P. Szolovits, Predicting icu mortality risk by grouping temporal trends from a multivariate panel of physiologic measurements, in: Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[4] L. A. Celi, S. Galvin, G. Davidzon, J. Lee, D. Scott, R. Mark, A database-driven decision support system: Customized mortality prediction, Journal of personalized medicine 2 (4) (2012) 138–148.

[5] R. Pirracchio, M. L. Petersen, M. Carone, M. R. Rigon, S. Chevret, M. J. van der Laan, Mortality prediction in intensive care units with the super icu learner algorithm (sicula): a population-based study, The Lancet Respiratory Medicine 3 (1) (2015) 42–52.

[6] V. J. Ribas, J. C. López, A. Ruiz-Sanmartín, J. C. Ruiz-Rodríguez, J. Rello, A. Wojdel, A. Vellido, Severe sepsis mortality prediction with relevance vector machines, in: Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE, IEEE, 2011, pp. 100–103.

[7] S. Kim, W. Kim, R. W. Park, A comparison of intensive care unit mortality prediction models through the use of data mining techniques, Healthcare informatics research 17 (4) (2011) 232–243.

[8] D. Delen, G. Walker, A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods, Artificial intelligence in medicine 34 (2) (2005) 113–127.

[9] E. D. Crawford, J. T. Batuello, P. Snow, E. J. Gamito, D. G. McLeod, A. W. Partin, N. Stone, J. Montie, R. Stock, J. Lynch, et al., The use of artificial intelligence technology to predict lymph node spread in men with clinically localized prostate carcinoma, Cancer 88 (9) (2000) 2105–2109.

[10] J. Calvert, Q. Mao, J. L. Hoffman, M. Jay, T. Desautels, H. Mohamadlou, U. Chettipally, R. Das, Using electronic health record collected clinical variables to predict medical intensive care unit mortality, Annals of Medicine and Surgery 11 (2016) 52–57.

[11] J. Lee, D. M. Maslove, Customization of a severity of illness score using local electronic medical record data, Journal of intensive care medicine 32 (1) (2017) 38–47.

[12] Á. Silva, P. Cortez, M. F. Santos, L. Gomes, J. Neves, Mortality assessment in intensive care units via adverse events using artificial neural networks, Artificial Intelligence in Medicine 36 (3) (2006) 223–234.

[13] W. A. Knaus, E. A. Draper, D. P. Wagner, J. E. Zimmerman, Apache ii: a severity of disease classification system., Critical care medicine 13 (10) (1985) 818–829.

[14] J.-R. Le Gall, S. Lemeshow, F. Saulnier, A new simplified acute physiology score (saps ii) based on a european/north american multicenter study, Jama 270 (24) (1993) 2957–2963.

[15] J.-L. Vincent, A. De Mendonça, F. Cantraine, R. Moreno, J. Takala, P. M. Suter, C. L. Sprung, F. Colardyn, S. Blecher, Use of the sofa score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study, Critical care medicine 26 (11) (1998) 1793–1800.

[16] S. Q. Simpson, New sepsis criteria: a change we should not make, CHEST Journal 149 (5) (2016) 1117–1118.

[17] G. B. Smith, D. R. Prytherch, P. Meredith, P. E. Schmidt, P. I. Featherstone, The ability of the national early warning score (news) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death, Resuscitation 84 (4) (2013) 465–470.

[18] J.-R. Le Gall, P. Loirat, A. Alperovitch, P. Glaser, C. Granthil, D. Mathieu, P. Mercier, R. Thomas, D. Villers, A simplified acute physiology score for icu patients, Critical care medicine 12 (11) (1984) 975–977.

[19] D. Poole, C. Rossi, N. Latronico, G. Rossi, S. Finazzi, G. Bertolini, et al., Comparison between saps ii and saps 3 in predicting hospital mortality in a cohort of 103 italian icus. is new always better?, Intensive care medicine 38 (8) (2012) 1280–1288.

[20] A. L. Rosenberg, Recent innovations in intensive care unit risk-prediction models, Current opinion in critical care 8 (4) (2002) 321–330.

[21] J.-L. Vincent, M. Singer, Critical care: advances and future perspectives, The Lancet 376 (9749) (2010) 1354–1361.

[22] M. T. Gilani, M. Razavi, A. M. Azad, et al., 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 (2) (2014) 144.

[23] W. A. Knaus, D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Bergner, P. G. Bastos, C. A. Sirio, D. J. Murphy, T. Lotring, A. Damiano, The apache iii prognostic system. risk prediction of hospital mortality for critically ill hospitalized adults., Chest Journal 100 (6) (1991) 1619–1636.

[24] J. R. Le Gall, A. Neumann, F. Hemery, J. P. Bleriot, J. P. Fulgencio, B. Garrigues, C. Gouzes, E. Lepage, P. Moine, D. Villers, Mortality prediction using saps ii: an update for french intensive care units, Critical Care 9 (6) (2005) R645.

[25] B. Metnitz, E. Schaden, R. Moreno, J.-R. Le Gall, P. Bauer, P. G. Metnitz, A. S. Group, et al., Austrian validation and customization of the saps 3 admission score, Intensive care medicine 35 (4) (2009) 616–622.

[26] M. Hoogendoorn, A. el Hassouni, K. Mok, M. Ghassemi, P. Szolovits, Prediction using patient comparison vs. modeling: A case study for mortality prediction, in: Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the, IEEE, 2016, pp. 2464–2467.

[27] J. E. Zimmerman, A. A. Kramer, D. S. McNair, F. M. Malila, Acute physiology and chronic health evaluation (apache) iv: hospital mortality assessment for todays critically ill patients, Critical care medicine 34 (5) (2006) 1297–1310.

[28] A. A. Kramer, T. L. Higgins, J. E. Zimmerman, Comparison of the mortality probability admission model iii, national quality forum, and acute physiology and chronic health evaluation iv hospital mortality models: implications for national benchmarking, Critical care medicine 42 (3) (2014) 544–553.

[29] T. L. Higgins, D. Teres, W. S. Copes, B. H. Nathanson, M. Stark, A. A. Kramer, Assessing contemporary intensive care unit outcome: an updated mortality probability admission model (mpm0-iii), Critical care medicine 35 (3) (2007) 827–835.

[30] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith, et al., The third international consensus definitions for sepsis and septic shock (sepsis-3), Jama 315 (8) (2016) 801–810.

[31] C. W. Seymour, V. X. Liu, T. J. Iwashyna, F. M. Brunkhorst, T. D. Rea, A. Scherag, G. Rubenfeld, J. M. Kahn, M. Shankar-Hari, M. Singer, et al., Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (sepsis-3), Jama 315 (8) (2016) 762–774.

[32] A. Awad, M. BaderElDen, J. McNicholas, Patient length of stay and mortality prediction: A survey, Health Services Management Research 30 (2) (2017) 105–120, pMID: 28539083. arXiv:https://doi.org/10.1177/0951484817696212, doi:10.1177/0951484817696212.
URL https://doi.org/10.1177/0951484817696212

[33] S. Lemeshow, D. Teres, H. Pastides, J. S. Avrunin, J. S. Steingrub, A method for predicting survival and mortality of icu patients using objectively derived weights., Critical care medicine 13 (7) (1985) 519–525.

[34] R. Dybowski, V. Gant, P. Weller, R. Chang, Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm, The Lancet 347 (9009) (1996) 1146–1150.

[35] A. Nimgaonkar, S. Sudarshan, Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models, Intensive Care Med 30 (2004) 248–253.

[36] G. Clermont, D. C. Angus, S. M. DiRusso, M. Griffin, W. T. Linde-Zwirble, Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models, Critical care medicine 29 (2) (2001) 291–296.

[37] L. Wong, J. Young, A comparison of icu mortality prediction using the apache ii scoring system and artificial neural networks, Anaesthesia 54 (11) (1999) 1048–1054.

[38] G. Doig, K. Inman, W. Sibbald, C. Martin, J. Robertson, Modeling mortality in the intensive care unit: comparing the performance of a back-propagation, associative-learning neural network with multivariate logistic regression., in: Proceedings of the Annual Symposium on Computer Application in Medical Care, American Medical Informatics Association, 1993, p. 361.

[39] L. Citi, R. Barbieri, Physionet 2012 challenge: predicting mortality of icu patients using a cascaded svm-glm paradigm, in: Computing in Cardiology (CinC), 2012, IEEE, 2012, pp. 257–260.

[40] J. Ramon, D. Fierens, F. Güiza, G. Meyfroidt, H. Blockeel, M. Bruynooghe, G. Van Den Berghe, Mining data from intensive care patients, Advanced Engineering Informatics 21 (3) (2007) 243–256.

[41] G. Meyfroidt, F. Güiza, J. Ramon, M. Bruynooghe, Machine learning techniques to examine large patient databases, Best Practice & Research Clinical Anaesthesiology 23 (1) (2009) 127–143.

[42] M. J. Berry, G. Linoff, Data mining techniques: for marketing, sales, and customer support, John Wiley & Sons, Inc., 1997.

[43] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The weka data mining software: an update, ACM SIGKDD explorations newsletter 11 (1) (2009) 10–18.

[44] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip, et al., Top 10 algorithms in data mining, Knowledge and information systems 14 (1) (2008) 1–37.

[45] M. Bader-El-Den, E. Teitei, M. Adda, Hierarchical classification for dealing with the class imbalance problem, in: Neural Networks (IJCNN), 2016 International Joint Conference on, IEEE, 2016, pp. 3584–3591.

[46] T. Perry, M. Bader-El-Den, S. Cooper, Imbalanced classification using genetically optimized cost sensitive classifiers, in: Evolutionary Computation (CEC), 2015 IEEE Congress on, IEEE, 2015, pp. 680–687.

[47] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000 (June 13)) e215–e220, circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

[48] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research (2002) 321–357.

[49] L. G. Glance, A. W. Dick, T. M. Osler, Icu scoring systems: After 30 years of reinventing the wheel, isnt it time to build the cart?, Critical care medicine 42 (3) (2014) 732–734.

[50] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, R. G. Mark, Mimic-iii, a freely accessible critical care database, Scientific data 3.