Universidad Nacional de Educación a Distancia (UNED).

Escuela Técnica Superior de Ingeniería Informática.

MÁSTER UNIVERSITARIO EN INTELIGENCIA ARTIFICIAL AVANZADA: FUNDAMENTOS, MÉTODOS Y APLICACIONES

Trabajo Fin de Máster "SISTEMAS INTELIGENTES DE DIAGNÓSTICO, PLANIFICACIÓN Y CONTROL":

Machine learning applied to a Cardiac Surgery Recovery Unit and to a Coronary Care Unit for mortality prediction

Autora: **Dr. Beatriz Nistal Nuño.**

Directores: Severino Fernández Galán, Pablo García Tahoces.

Madrid, Junio 2021.

Contents:

ABSTRACT

Predicting the mortality risk for patients with cardiac disease in Intensive Care Units is essential for effective care planning, where impending deterioration can occur with severe health consequences. Most established severity-of-illness systems used for prediction of Intensive Care Unit mortality were developed targeted at the general Intensive Care Unit population, based on logistic regression. To date, no dynamic predictive tool has been developed targeted at patients in the Cardiac Surgery Recovery Unit and Coronary Care Unit using machine learning. In this research, adult patients at the Cardiac Surgery Recovery Unit and Coronary Care Unit from the MIMIC-III critical care database were studied. Intensive Care Unit data was extracted during a 5 hour window in addition to a few demographic features to produce 12 hour advance mortality predictions. The machine learning models developed were the Tree Ensemble of Decision Trees, Random Forest of Decision Trees, XGBoost Tree Ensemble, Naive Bayes, and Bayesian network. The models were compared to six established systems by assessing the discrimination, calibration and accuracy statistics. The main advantages of these models are that they overcome most limitations of logistic regression, utilized to build the established systems, in addition to being dynamic as opposed to the static traditional systems. The AUROC values for the primary outcome were superior for all the machine learning models, the accuracy statistics less sensitive to unbalanced cohorts were substantially higher for all the machine learning models and, finally, the Brier score was better for all the machine learning models except the Naïve Bayes over the conventional systems. In conclusion, the discriminatory power of XGBoost and Tree Ensemble were excellent, substantially outperforming the conventional systems. Additionally, the machine learning models showed better performance for the vast majority of accuracy measures. Consequently, the models developed in this work offer promising results that could benefit Cardiac Surgery and Coronary Care Units.

Keywords: Cardiac Surgery, Coronary Care Unit, machine learning, mortality, discrimination, calibration.

1. INTRODUCTION

Prediction of mortality risk for patients with cardiac disease in both medical and cardiac surgery intensive care units (ICU) is needed for effective care planning and appropriate management of these kind of patients, where impending deterioration can be quick and with severe health consequences or irreversible. The acute pathophysiological consequences of cardiopulmonary bypass in patients experiencing cardiac surgery are temporary and physiologic alterations can be undetected by support devices, like intra-aortic balloon pumps, mechanical circulatory support devices such as ventricular assist devices, extracorporeal membrane oxygenation, renal replacement therapy and mechanical ventilation. These support devices are also commonly used by the population of patients in the Coronary care unit (CCU). In addition, the effects of the Anesthesia strategy during cardiac surgery can have an impact on the short-term and long-term survival after cardiac surgery [1]. The patients in the Cardiac Surgery Recovery Unit (CSRU) and CCU comprise a heterogeneous population with varied multiorgan dysfunctions with predominant cardiac disease, very different in comparison to other surgical or medical ICU patients. These characteristics make the subset of patients with cardiac disease,

in the CSRU and CCU, a particular subgroup where predictor tools targeted to this population would be beneficial.

Several severity-of-illness scoring systems for mortality risk prediction in the ICU have been developed for general ICU patients, and are currently broadly used in ICUs including the settings of the CSRU and CCU. These are for example the sequential organ failure assessment score (SOFA), the Simplified Acute Physiology Score (SAPS), the SAPS II, the SAPS III, the Logistic Organ Dysfunction System (LODS), and the Oxford Acute Severity of Illness Score (OASIS). Additionally, diagnosis-specific risk scores have been validated for ICU patients for predicting mortality in specific subsets of patients. Regarding the population of adult cardiac patients, two scores have been developed to date to estimate the risk of mortality in the ICU. One is the Mayo CICU Admission Risk Score (M-CARS), for mortality risk assessment of patients in the CCU. The other is the Cardiac Surgery Score (CASUS) and its logistic version, the Log-CASUS, which was designed for patients admitted postoperatively to the ICU after cardiac surgery. Logistic regression (LR) has been used to construct the majority of the mentioned prediction tools and continues to be a popular method. There are several reasons why machine learning (ML) methods can render a more accurate prediction of ICU mortality, which will be noted in the next section.

The majority of studies applying ML for the prediction of ICU mortality use static models collecting the initial data of ICU admission to build the model. The creation of real-time predictors is important in order to follow over time the events that may change the prognosis for an individual patient during their ICU stay. There are some examples with regard to dynamic models of adult mortality prediction in the general ICU built by ML methods. Other studies have applied ML to predict ICU mortality for specific subsets of ICU patients. Some of these studies will be outlined in the next section.

To date, no predictive tool for ICU mortality has been developed targeted at unselected adult patients both in the CSRU and CCU using ML techniques, being a dynamic predictor producing an updated individual ICU mortality risk estimation during the ICU stay. Many of the established severity-of-illness scoring systems are used in the CSRU and CCU, but they were designed for general ICU patients instead. Additionally, these conventional systems are static as they are calculated from values obtained during the first 24 hours of the ICU stay and were thus not designed to track a patient´s health status evolution during the ICU stay, providing an initial risk stratification at the time of ICU admission. It is investigated in the current work the predictive performance of five ML methods for 12-hour ICU mortality for individual adult patients in the CSRU and CCU. Additionally, it is studied whether they outperform the prognostic performance of six conventional severity-of-illness scoring systems.

The rest of this paper is organized as follows. Section 2 reviews conventional methods developed for mortality risk prediction in the ICU and some recent examples of ML methods developed for this outcome. Section 3 explains in detail the characteristics of the methods proposed in this work for mortality prediction in the ICU and in-hospital. Section 4 evaluates the proposed methods and analyzes the experimental results obtained. Finally, Section 5 enumerates the main contributions of the present work and the most promising research directions derived from it.

2. RELATED WORK

This section first describes traditional scoring systems widely established in the ICU for mortality prediction and other scoring systems developed using also conventional methods but that are not so

widely established. Next, some recent studies that use ML techniques for ICU mortality prediction are outlined.

2.1. CONVENTIONAL METHODS FOR MORTALITY RISK PREDICTION IN THE INTENSIVE CARE UNIT

Conventional prognostic scoring systems for mortality risk prediction for general ICU patients have been developed, and are currently broadly used both in the CSRU and CCU. These comprise the SOFA, which is used to follow a patient's evolution during the ICU stay to assess the degree of a patient's organ function [2,3]. The SAPS was developed as a simplification of the Acute Physiology Score (APS), reducing the number of required parameters [4]. The SAPS II was designed to evaluate the severity of disease for ICU patients of 15 years old or more. The measurement is completed 24 hours after ICU admission providing an integer score from 0 to 163, from which the mortality risk is estimated with calibration coefficients [5]. The posterior modification, the SAPS III, attempts to account for poor calibration of SAPS II. LR was used to compute the probability of mortality [6]. The SAPS III data was not illustrative of all kinds of patient populations as it was created using a general ICU population. Although the subset of CSRU patients was originally excluded during the development of the SAPS, it is used nowadays in cardiac ICUs [7].

In the LODS, physiological variables indicate dysfunction in six organ systems. LR was used to estimate severity levels and relative weights for the score and for conversion of the LODS score into a probability of mortality [8]. The OASIS was developed by Johnson et al. ML algorithms of type *particle swarm optimization* were used to select the minimal set of variables that were capable of yielding an accurate severity-of-illness score [9].

Diagnosis-specific risk scores have been validated for ICU patients. For example, two scores were specifically developed for cardiac arrest patients [10]. One is the multivariate LR model out-of-hospital cardiac arrest (OHCA) score, developed by Adrie et al. They used LR to discover clinical and laboratory variables readily available at ICU admission predictive of mortality and neurological outcomes [11]. The Cardiac Arrest Hospital Prognosis (CAHP) score was developed by Maupain et al. Independent prognostic factors were identified using LR analysis. The CAHP score is a simple system for early stratification of patients admitted in ICU after OHCA [12]. Studies have demonstrated the superiority of OHCA and CAHP over general severity-of-illness scores [10,13] for predicting mortality in cardiac arrest patients.

For mortality risk assessment in general adult cardiac patients in the CCU, Jentzer et al. evaluated the ability of the SOFA score to predict mortality in a cohort of 9961 unselected adult patients in the CCU. The day 1 SOFA score had good discrimination for short-term mortality, similar to Acute Physiology And Chronic Health Evaluation (APACHE)-III and IV [14]. Posteriorly, Jentzer et al. designed a score as a tool to provide rapid mortality risk stratification at the time of CCU admission to facilitate patient triage and recognition of high-risk cardiac patients, using parameters available at the time of CCU admission [15]. The best 7 predictors of hospital mortality were found using stepwise backward regression, and utilized to create the M-CARS system [15].

Regarding the population of general adult patients in the CSRU, the CASUS was developed by Hekmat et al. as a system for daily risk stratification of patients admitted postoperatively to the ICU after cardiac surgery [16]. This score was a simple additive model, with its points allocated for dysfunction of 6 organ systems providing a total number which estimates the risk of mortality. It was compared to

general severity-of-illness systems for ICU mortality prediction after adult cardiac surgery in a posterior study by Doerr et al. [17]. They concluded that CASUS and SOFA are reliable ICU mortality risk stratification models for cardiac surgery patients, while the SAPS II and APACHE II showed worse results [17]. Doerr et al. presented the logistic version of the additive CASUS (Log-CASUS) in 2012. The logistic model showed statistical superiority [18].

Linear models such as LR analysis have been used commonly to construct such prognostication tools and continue to be a popular method for constructing new risk predictor tools for varied outcomes. LR is a statistical algorithm that models the relationship between the input features and the categorical output classes by maximizing a likelihood function. In its basic form uses a logistic function to model a binary dependent variable. ICU prediction models such as the APACHE III and IV, the Mortality Probability Model, SAPS II, SAPS III and LODS are based on multivariable LR models. The conventional widely used systems were universally agreed to lack sufficient calibration to be used on the individual level [19]. SAPS III calculated in patients admitted to ICU after cardiac arrest showed only moderate discrimination [13]. Some bases for the low predictability of many of these scoring tools are the assumptions of patient features' independence and that the features are linearly related to the log odds of outcomes when using LR statistical methods, revealing insufficient performance. There is increasing evidence that ML models can render a more accurate outcome prediction for ICU patients to support decision making. ML models can yield new insights into complex interactions, non-linearities, and the significance of trends in the feature variables.

2.2. RECENT MACHINE LEARNING METHODS FOR MORTALITY RISK PREDICTION IN THE INTENSIVE CARE UNIT

The majority of studies applying ML for the prediction of ICU mortality use static models collecting the initial data of ICU admission to build the model, with the purpose of early mortality risk stratification or triage. The creation of real-time predictors is important for not missing the events that distinctly influence the prognosis for an individual patient during their ICU stay. There are some examples with regard to dynamic models of adult mortality prediction in the general ICU. The study of Johnson et al. evaluated an ML model comparing it to established severity-of-illness systems building a real-time mortality prediction tool. The model used was Gradient Boosting Decision Trees (GB) compared to several types of LR. The GB model greatly outperformed the SAPS II [19]. The study of Thorsen-Meyer et al. investigated the real-time prediction of 90-day mortality for adult patients in the ICU using a deep learning model. Their ML model showed a predictive performance that improved along the time course of an ICU stay and revealed good calibration [20].

Other studies have applied ML to predict ICU mortality for specific subsets of ICU patients, like the study of Jain et al. that constructed different predictive models including ML to assess mortality during the ICU stay for patients admitted with acute exacerbation of chronic obstructive pulmonary disease [21]. Nanayakkara et al. used ICU data available within the first 24 hours of ICU stay to develop more accurate models of risk prediction of in-hospital mortality for adult cardiac arrest patients. They used LR and ML techniques to compare to the APACHE III and to the Australian and New Zealand Risk of Death predictions. ML models significantly improved the discrimination and calibration [22].

3. MACHINE LEARNING METHODS USED IN THIS WORK

The models developed in this work have the potential to follow over time the health status evolution of individual patients both in the CSRU and CCU, in comparison to the widely established general

ICU scoring systems that are calculated only once (shortly after ICU admission). Additionally, these models use only 5 consecutive hours of data in comparison to the 24 hours of data required by such general ICU scoring systems in order to predict ICU mortality. These methods use ML in the form of ensemble learning of decision trees (DTs) and probabilistic models such as Bayesian network (BN) and Naive Bayes (NB) in order to compose more advanced models overcoming most limitations of LR in this clinical setting, used commonly to construct these prediction tools.

3.1. PATIENT COHORT AND DATA EXTRACTION

The Medical Information Mart for Intensive Care (MIMIC) III critical care database version v1.4. was used, a large database comprising de-identified clinical data of individual patients admitted to ICUs between 2001 and 2012 at the tertiary care hospital Beth Israel Deaconess Medical Center (BIDMC) in Boston (USA). The project for this database creation was approved by the Institutional Review Boards of the BIDMC and the Massachusetts Institute of Technology (Cambridge, USA) [23,24]. The data of the MIMIC-III database was installed in a local PostreSQL database management system on Linux.

A final cohort of 11059 ICU-stay patient records from the MIMIC III database was used in this work, which were selected according to the following data extraction steps. The selected subgroup consisted of the ICU-stay records of adult patients of age 16 years old or more admitted to the CSRU and CCU, and with a documented length-of-stay and survival for at least 36 hours following ICU admission. The cutoff of 36-h length-of-stay was chosen to allow 24-h of data collection in the ICU for calculation of all the severity-of-illness scores used for comparison, as the calculation window for these systems is 24 hours (except for the SAPS III which is 1 h), in order to avoid using these scores before the end of the first ICU day. The ML methods developed in this work extract data during a 5-h window in addition to a few demographic features to produce 12-h advance predictions. The cutoff of 36 hours allows then to generate an alert at 12 h prior to patient death or discharge from the ICU for the ML methods developed and to compare to the conventional systems. The ML methods are also compared directly to a serial SOFA score which is calculated also at 12-h prior to patient death or discharge in parallel to the ML methods, based on 24-h of data, as this system has been established to be used serially and updated during the ICU stay [14,17].

Out of the final 11059 ICU-stay patient records selected, 721 resulted in ICU death during the ICU stay. That amounts to a prevalence of 6.52% of ICU mortality. In-hospital mortality occurred for 959 ICU stays during the same hospital admission of the corresponding ICU stays, amounting to a prevalence of 8.67% for in-hospital mortality. This cohort was used for all experiments. The primary outcome prediction endpoint was all-cause mortality during the ICU admission, and the secondary outcome prediction endpoint was all-cause in-hospital mortality during the same hospital admissions of the corresponding ICU stays. For the primary and secondary outcomes, the ML models were compared to the severity-of-illness scores of OASIS, SAPS, SAPS II, SAPS III, LODS and SOFA. These established systems´ scores were generated using open source code accompanying the MIMIC III database [25].

3.2. FEATURE SELECTION

Both dynamic and static data was obtained from the final cohort. The 1-h time-resolution physiological and laboratory measurements collected during 5 consecutive hours from the 11059 ICU-stay records consisted of heart rate, pH, pulse pressure, respiration rate (RR), blood oxygen saturation (So2), systolic blood pressure (SBP), mean blood pressure, temperature, hemoglobin and white blood cell

count (WBC). These variables were selected because they are relevant measurements frequently available in the ICU. Variables measured more than once per hour were down-sampled, yielding the first hourly value. As laboratory values like pH, Hemoglobin, and WBC were measured less frequently than hourly, their window for extraction was extended 24 hours backward for their first hourly measurement of the 5-h period. These values were of type double for the different measurements.

Static features were collected as demographic variables at the time of ICU admission including type of ICU admission, gender, insurance status, ethnicity, and age at ICU admission. The type of ICU admission was categorized as elective, urgent, or emergency. There were seven primary ethnicities designated in MIMIC-III. To measure the impact of socioeconomic status, insurance status as categorized by MIMIC-III was used: Medicare, Private, Medicaid, Government, or Self Pay.

3.3. DATA PRE-PROCESSING

Data pre-processing was performed based on domain knowledge to delete erroneous recordings such as physiologically spurious entries and unit conversion errors. For a single missing hourly value, a replacement was calculated as the available value immediately preceding during the 5-h window. For a missing value in the first hourly measurement of the 5-h window, a replacement was calculated as the average of values for the entire cohort for that parameter at that time point.

The input data-set was split randomly into two partitions, 80% for training data and 20% for testing data (see Figure 1). The ML models were built with the training data and their performance was evaluated on the testing data. The training data was composed of 8847 ICU-stay records and the testing data was composed of 2212 ICU-stay records. The same testing data (n=2212) was used to evaluate the performance of all ML models developed and that of all the severity-of-illness systems compared. The scores of these conventional systems were calculated on this same testing data [25].

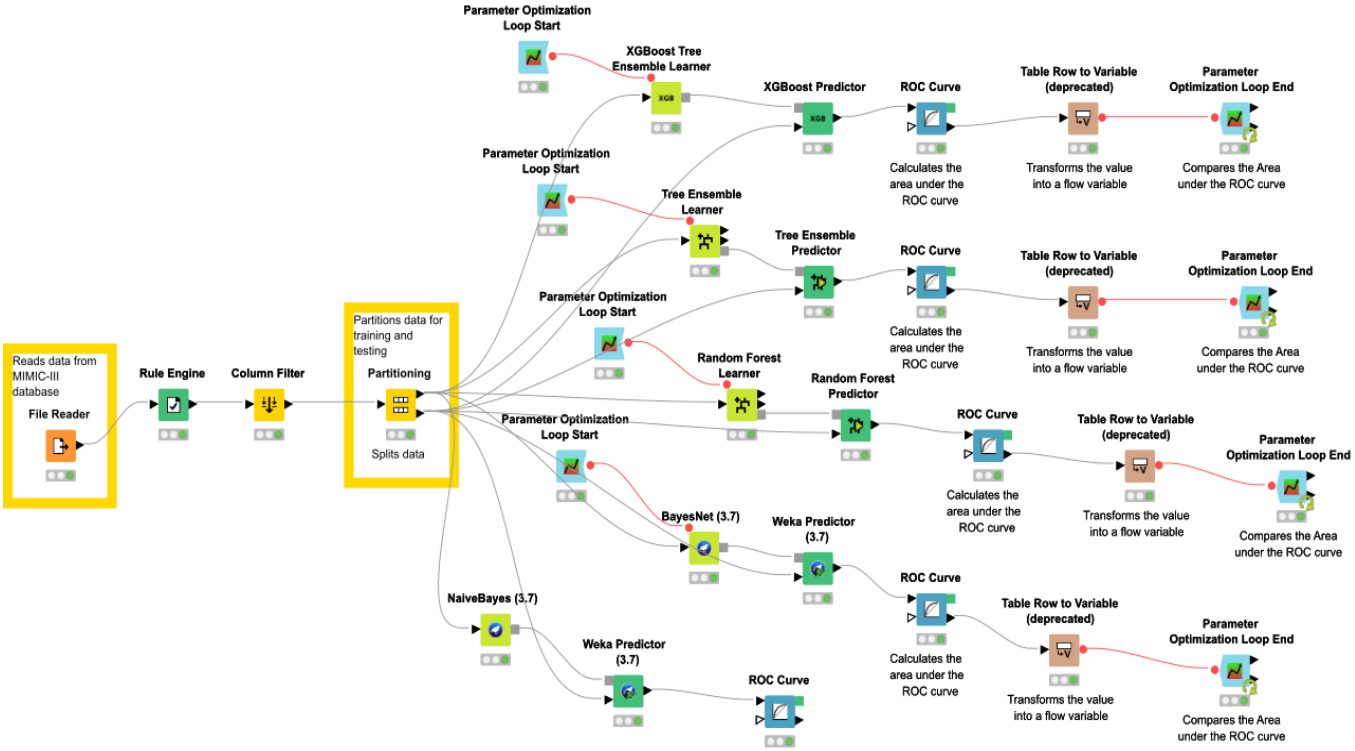


Figure 1. KNIME workflow. Screen-shot of the KNIME workflow used to build the Machine Learning models.

3.4. MACHINE LEARNING MODELS

Supervised learning on labeled data was used. Among the supervised learning algorithms, the ones used to construct the ML models in this work were: (a) Probabilistic models like NB and BN, and (b) Ensemble learning models like the Tree Ensemble of DTs (TE), Random Forest of DTs (RF), and the XGBoost Tree Ensemble (XGB). The data was imported into KNIME *(KNIME AG, Zurich, Switzerland)* version 4.3.0 to execute the simulations, where all the models were implemented [26].

The two-class target variable was *icustay_expire_flag* or *hospital_expire_flag* for the primary and secondary outcomes respectively.

3.4.1. Probabilistic models

The model learns to estimate a class probability and to assign each vector of input features to the class with the highest probability [26].

3.4.1.1. Naive Bayes

It is based on the Bayes' theorem and assumes statistical independence between the input features given the output class ("naive"). This algorithm estimates the conditional probability of each output class given the vector of input features. The class with the highest conditional probability is assigned to the input data [26].

It was implemented with the NaiveBayes (3.7) node and the Weka_Predictor (3.7) node, based on WEKA 3.7. The NaiveBayes (3.7) node creates an NB model from the training data. It calculates a Gaussian distribution per class for the numerical attributes used [26]. The created model is used in the predictor node, which predicts the class of unclassified test ICU-records at the inport (see Figure 1).

This NB classifier has the following parameters:

1. Kernel Estimator: used normal distribution instead of kernel density estimator for numeric attributes.

2. Supervised Discretization: uses supervised discretization to convert numeric features to nominal ones [26].

3.4.1.2. Bayesian network

BNs [27,28] can construct expert systems by employing a statistical design that provides a graphical description of the probabilistic connections between a group of variables. This BN was composed of the observed variables (the patient features) and the interest variable *icustay_expire_flag* or *hospital_expire_flag* (the variable whose a posteriori probability is of interest) in an acyclic directed graph. In this graph, each node depicts a variable connected by links defined over the nodes according to a probability distribution over the variables.

It was implemented with the BayesNet (3.7) node and the Weka_Predictor (3.7) node, based on WEKA 3.7 [26]. The BayesNet node creates the network structure and finds the conditional probability distributions. The Weka_Predictor node takes this model and classifies the test ICU-records at the inport (see Figure 1).

This BN classifier has the following parameters, of which 3 and 4 were optimized through the Parameter_Optimization loop (see Section 3.5):

1. Search Algorithm: the BN learning algorithm, the method used for searching the network structure learned from data. The K2 algorithm was used, which uses a hill climbing strategy restricted by variable order. The order of the variables in the training data was used for the order of the nodes in the network. As the network was initialized as NB network for structure learning, the outcome variable was first in the ordering. The score type used to judge the quality of the network structure was Bayes [26].
2. Estimator: the algorithm for discovering the conditional probability tables of the BN once the structure has been learned [26].
3. Alpha: is used for estimating the probability tables and can be explained as the initial count on each value [26].
4. Maximum number of parents: the maximum number of parents a variable could have in the BN [26].

3.4.2. Ensemble learning

It entails a combination of multiple models from supervised learning algorithms to achieve a more stable and accurate overall model. The ensemble methods used were Bagging and Boosting [26].

3.4.2.1. Bagging

A method for training multiple classification models on different randomly drawn subgroups of the training dataset. The final prediction is grounded on the predictions provided by all the models, thus reducing the chance of over-fitting [26].

3.4.2.1.1. Tree Ensemble of Decision Trees

A tree ensemble of DTs corresponds to an ensemble model of multiple DTs trained on different subsets of data. Data subsets with less or equal records and less or equal features are bootstrapped from the original training set.

A DT is a representation for classifying data. Each non-root node of the DT is represented with a discrete input feature. The root of the tree is represented with a probability distribution over the classes that represent the outcome. The underlying algorithm for constructing a DT performs a recursive binary partitioning of the feature space. The best split is selected from a set of possible splits, in order to maximize the information gain or Gini index at a tree node [26].

It was implemented by the Tree_Ensemble_Learner node and the Tree_Ensemble_Predictor node (see Figure 1) [26]. This model learns an ensemble of DTs. Each of the DTs is learned on a different group of ICU records and/or a different group of patient features. The output model is applied in the corresponding predictor node which yields the final prediction of mortality according to the aggregation mode of "hard voting" (majority rule) to aggregate the predictions of the individual DTs. The class that obtains the most votes is selected [29].

This TE has the following parameters, of which 1, 2, 3 and 5 were optimized through the Parameter_Optimization loop (see Section 3.5):

1. Tree depth: maximum number of tree levels to be constructed.
2. Number of models: the number of DTs to learn.
3. Fraction of Data Sampling: the sampling of the ICU-records for learning each individual DT.
4. Data Sampling Mode: it determines how the ICU-records from the training data set are sampled. It was random with replacement.
5. Attribute Sampling: indicates the sampling of the features to learn a single DT. This was a function (linear fraction) of the number of features.
6. Attribute Selection: It sampled a different group of candidate features in each of the tree nodes from which the optimal one was chosen to execute the split [26].

3.4.2.1.2. Random Forest of Decision Trees

It was implemented by the Random_Forest_Learner node and the Random_Forest_Predictor node (see Figure 1) [26]. This model provides a subset of the functionality of the TE. The RF_Learner node learns a RF, which consists of a chosen number of DTs. Each of the DTs is learned on a different group of patient features. The groups of ICU-records for each DT are the same as the original training set. In order to perform the split at each node of a DT, a new group of patient features is selected by taking a random sample of size *sqrt(m)* where *m* is the original total number of patient features. The output model is applied in the corresponding predictor node for the final prediction, which is based on a majority rule on all involved DTs (see section 3.4.2.1.1) [29].

The main differences of RF in comparison to the TE model are:

1. In the TE, a single DT can be learned on a fraction of ICU-records or the whole data like the RF, having different options for modes of data sampling.

2. In the TE, feature sampling to learn the DTs can occur or not (using all features), and there are several options for types of feature sampling instead of only one for the RF.
3. In the TE, the feature selection applied when features are sampled can happen at the DT level or at the tree node level, while only per tree node for the RF.
4. A fixed root feature can be chosen for the TE, used in all DTs even if the feature is not in the feature sample [26].

This RF has the following parameters, which were optimized through the Parameter_Optimization loop (see Section 3.5):

1. Tree depth: maximum number of tree levels to be constructed.
2. Minimum node size: minimum number of ICU records in child nodes. It can be at most half of the minimum split node size.
3. Number of models: the number of DTs to learn [26].

3.4.2.2. Boosting

A method for training a group of classification models iteratively. At each iteration, a new model is trained on the prediction errors and added to the ensemble to improve the results from the previous model state, providing higher performance after each iteration [26].

3.4.2.2.1. XGBoost Tree Ensemble

Extreme Gradient Boosting (XGBoost) belongs to the gradient boosting framework. The XGB learns a tree based XGBoost model for classification. XGboost makes use of a gradient descent algorithm for optimization, improving the predictive performance at each optimization step by following the negative of the gradient as it is trying to find the "sink" in an *n*-dimensional plane. XGB minimizes a regularized objective function that merges a convex loss function, which is based on the variation between the target outputs and the predicted outputs. Combining the loss function with a regularization term arrives at the objective function. The regularization term controls the complexity and reduces the risk of over-fitting. The training adds new trees at each iteration, with the capability to predict the residuals as well as errors of prior trees, which are then coupled with the previous trees to calculate the final prediction [26].

It was implemented by the XGBoost_Tree _Ensemble_Learner node and the XGBoost_Predictor node (see Figure 1) [26].

This XGB has the following parameters, of which all except 2, 3, and 14 were optimized through the Parameter_Optimization loop (see Section 3.5):

1. Boosting rounds: the number of models to train in the ensemble.
2. Objective: it was used the softprob objective function.
3. Booster: the default tree booster was used.
4. Eta: the learning rate. Step size shrinkage used in updates so as to avoid overfitting. The boosting process is more conservative as the Eta value decreases [26].
5. Gamma: minimum loss reduction needed to make an additional partition on a leaf node. The algorithm is more conservative as the Gamma increases.
6. Maximum depth of a DT.
7. Minimum child weight: minimum sum of instance weight (Hessian) required in a child. If the DT

partition step results in a leaf node with the sum of instance weight less than minimum child weight, then the construction will stop. The larger minimum child weight is, the more shallow the DT will be [26].

8. Maximum delta step: maximum delta step we permit every leaf output to be. The update step will be more conservative if this is set to a positive value [26].
9. Subsample ratio of the training instances: it can help to decrease over-fitting. Subsampling will happen once in every iteration.
10. Feature sampling rate by tree: Subsample ratio of features when building each DT. Subsampling will happen once in every iteration.
11. Feature sampling rate by level: Subsample ratio of features for each split.
12. Lambda: L2 regularization term on leaf weights. The model will be more conservative as this value increases.
13. Alpha: L1 regularization term on leaf weights. The model will be more conservative as this value increases.
14. Tree method: the DT building algorithm used. It was Auto, which uses a heuristic to select the fastest method.
15. Scale positive weight: manages the balance of positive and negative weights, convenient for unbalanced cohorts [26].

3.5. PARAMETER OPTIMIZATION

It was utilized the technique of Parameter Optimization with a parameter optimization loop in order to find the optimal parameters for the different ML models, trying to find an optimal design for each one of the ML models. It was implemented with the Parameter_Optimization_Loop_Start node and the Parameter_Optimization_Loop_End node (see Figure 1) [26].

The parameters described above, controlled via flow variables, were chosen by an algorithm to maximize the area under the receiver operator characteristic (ROC) curve (AUROC) (see Section 3.6) during the simulations for the corresponding outcome. The loop varies the several parameters following the search strategy of Random Search with a predefined maximum number of iterations: parameter combinations are randomly selected with replacement and assessed. As early stopping was used, the search stops when the objective value does not improve for a specified number of rounds [26].

The best parameter values found during the loops after several optimization simulations for the primary outcome for each one of the ML models are shown in Table 1. The remainder parameters were set to their default values.

Table 1. The best parameters found during the Parameter Optimization loop for the primary outcome for each one of the Machine Learning models developed.

	BN	**NB**	**XGB**	**TE**	**RF**
Estimator[a]	Simple Estimator				
Search algorithm[a]	K2				
Use Kernel Estimator[a]		False			
Use Supervised Discretization[a]		True			
Boosting rounds			60		
Eta			0.3		
Gamma			1		
Maximum depth (levels) of a tree			29	47	57
Minimum child weight			0.5		
Maximum delta step			5		
Subsample ratio of the training instances			0.8	0.8	
Feature sampling rate by tree			1		
Feature sampling rate by level			0.8	0.2	
Lambda			0.7		
Alpha for estimating the probability tables	0.982				
Alpha (L1 regularization term on leaf weights)			0.9		
Maximum number of parents	43				
Scale positive weight			12.7		
Number of models (DTs)				620	655
Minimum node size					110

[a]These parameters are network options that are set before the training optimization simulations.

DT =decision tree; NB=Naïve Bayes; BN=Bayes Network; TE=Tree Ensemble; RF=Random Forest; XGB=XGBoost Tree Ensemble.

3.6. PERFORMANCE MEASURES

The AUROC (ideal value: 1) was used to evaluate the discrimination performance of the models, together with the ROC curves. The ROC curve is a graphical plot that illustrates the discrimination ability of a binary classifier as its diagnostic threshold is varied, plotting the true positive rate or sensitivity against the false positive rate or (1 − specificity). The Empirical ROC curve was the one used [30].

The AUROC values were calculated for all models for the primary and secondary outcomes. They were based on the scores calculated from the first 24 hours of ICU stay for the conventional systems and based on 5 hours of ICU data for the ML models. It was also calculated the AUROC for the serial SOFA, whose scores are calculated in parallel to the ML models from the 24 hours preceding the 12-hour advance prediction prior to death or discharge from the ICU. The models were also evaluated

using the precision recall curves (PRC) for the primary and secondary outcomes, which provide a measure of performance which is agnostic to the number of true negatives and can be convenient for cohorts with class imbalance [31].

The metric for evaluating the calibration of the different predictive models was the Brier score. This was calculated for the primary and secondary outcomes for all the ML models and the OASIS, SAPS II and SAPS III. It was calculated as the mean squared error of the forecast [32].

Several accuracy statistics were used to evaluate and compare the ML models and conventional systems: sensitivity or recall, specificity, positive predictive value (PPV) or precision, negative predictive value (NPV), diagnostic odds ratio (DOR), overall Accuracy, Cohen's kappa (CK), F-measure, Matthews correlation coefficient (MCC), Balanced accuracy (BACC), and Markedness (MK) [32]. These measures were calculated for the primary outcome for all models.

Sensitivity is the ability of the predictor to correctly identify patients that are positive. Specificity is the ability of the predictor to correctly identify patients that are negative. PPV is the probability that patients with a positive prediction are truly positive. NPV is the probability that patients with a negative prediction are truly negative. The DOR is defined as the ratio of the odds of the prediction being positive if the patient is positive relative to the odds of the prediction being positive if the patient is negative. The F-measure is defined as the weighted harmonic mean of the precision and recall of the test. The CK takes into account the a priori distribution of the target classes, accounting for the chance of random classification, with values of 1 suggesting a perfect classification. The MCC can be thought of as a discretization of the Pearson correlation for binary variables, with +1 representing a perfect prediction. The MK is a measure of trustworthiness of positive and negative predictions by a system. Balanced accuracy is calculated as the average of the proportion corrects of each class individually. The DOR, CK, F-measure, MCC, BACC, and MK are considered useful for unbalanced data-sets as in this cohort [32].

The same testing data (n=2212) was used to calculate all these performance measures.

4. RESULTS

Figure 2 shows the ROC curves for all the ML models developed and three conventional systems (OASIS, SAPS II, and SAPS III) evaluated for ICU mortality prediction in the CSRU and CCU, displaying the AUROC values for these models. The AUROCs for all models for ICU mortality prediction are shown in Table 2. The serial SOFA score at a same time point 12 hours before ICU death or discharge yielded the AUROC of 0.8405 for ICU mortality prediction (Table 2).

Figure 3 shows the ROC curves for all the ML models and three conventional systems for the secondary outcome, displaying the AUROC values for these models. The AUROCs for all models for in-hospital mortality prediction are shown in Table 2. The serial SOFA score at a same time point 12 hours before ICU death or discharge yielded the AUROC of 0.8093 for in-hospital mortality prediction (Table 2).

Figure 4 shows the PRC for the SAPS II, OASIS, SAPS III and all the ML models for mortality prediction in the CSRU and CCU. Figure 5 shows the PRC for the SAPS II, OASIS, SAPS III, and all the ML models for in-hospital mortality prediction.

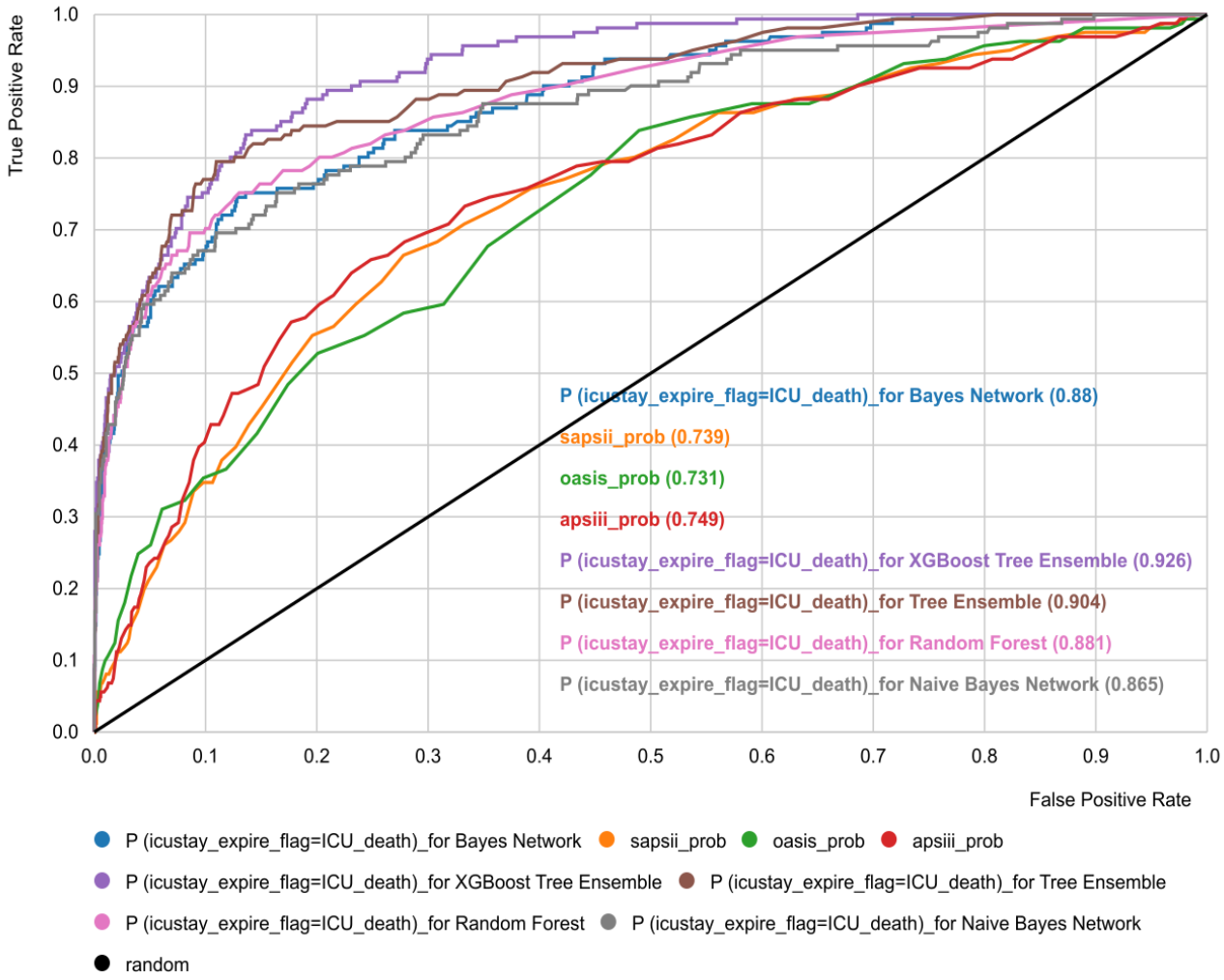


Figure 2. Receiver Operating Characteristic curves for mortality prediction in the Cardiac Surgery Recovery Unit and Coronary Care Unit for the OASIS, SAPS II, SAPS III, and the Machine Learning models developed. The two-class target variable was *icustay_expire_flag.*

Table 2. Comparison of the different Machine Learning models´ performance with the established severity-of-illness scoring systems for the prediction of mortality in the Cardiac Surgery Recovery Unit, Coronary Care Unit, and in-hospital[a].

	range	NB	BN	TE	RF	XGB	OASIS	SAPS III	SAPS II	LODS	SOFA	Serial SOFA	SAPS
Threshold		0.112	0.0411	0.1113	0.03992	0.046	36	59	44	8	6	5	21
AUROC for mortality in CSRU and CCU	[0,1]	0.865	0.88	**0.904**	0.881	**0.926**	0.731	0.749	0.739	0.6903	0.7068	**0.8405**	0.6851
AUROC for in-hospital mortality	[0,1]	0.797	**0.835**	0.827	0.802	**0.858**	0.72	0.754	0.742	0.6855	0.6771	**0.8093**	0.6716
Sensitivity	[0,1]	0.696	0.72	0.801	0.733	0.795	0.596	0.547	0.627	0.36	0.64	0.7763	0.64
Specificity	[0,1]	0.885	0.877	0.874	0.881	0.881	0.686	0.834	0.742	0.868	0.674	0.7893	0.61
PPV	[0,1]	0.322	0.315	**0.332**	0.326	**0.344**	0.13	0.205	0.16	0.176	0.133	0.2244	0.114
NPV	[0,1]	0.974	0.976	**0.982**	0.977	**0.982**	0.956	0.959	0.962	0.945	0.96	0.9782	0.956
DOR	[0,∞]	17.579	18.402	**27.892**	20.323	**28.725**	3.227	6.045	4.843	3.699	3.669	13.012	2.783
Accuracy	[0,1]	0.871	0.866	0.868	0.87	**0.875**	0.679	0.813	0.734	0.831	0.671	0.7884	0.613
CK	[-1,1]	0.378	0.375	**0.409**	0.39	**0.422**	0.106	0.215	0.158	0.154	0.114	0.2652	0.08
F-measure	[0,1]	0.44	0.439	**0.47**	0.451	**0.48**	0.213	0.298	0.255	0.237	0.221	0.3481	0.194
MCC	[-1,1]	0.414	0.417	**0.461**	0.431	**0.47**	0.155	0.25	0.213	0.167	0.171	0.3386	0.132
Brier score	[0,1]	0.084	0.046	**0.043**	0.047	**0.041**	0.075	0.068	0.12				
Brier score for in-hospital mortality	[0,1]	0.107	**0.061**	0.064	0.066	**0.059**	0.088	0.082	0.124				
BACC	[0,1]	0.79	0.799	**0.837**	0.807	**0.838**	0.641	0.69	0.685	0.614	0.657	0.7828	0.625
MK	[-1,1]	0.296	0.291	**0.315**	0.303	**0.326**	0.086	0.164	0.122	0.122	0.093	0.2026	0.07

[a]Results presented are based on test set (*n*=2212).

PPV=positive predictive value; NPV=negative predictive value; DOR=diagnostic odds ratio; BACC=balanced accuracy; MK=Markedness; MCC=Matthews correlation coefficient; CK=Cohen's kappa; CSRU=Cardiac Surgery Recovery Unit; CCU=Coronary Care Unit; AUROC=area under the receiver operator characteristic curve; NB=Naïve Bayes; BN=Bayes Network; TE=Tree Ensemble; RF=Random Forest; XGB=XGBoost Tree Ensemble.

Table 2 shows the accuracy statistics for the ML models and the conventional systems. The thresholds shown are the ones used to generate the tabulated accuracy statistics for the prediction of mortality in CSRU and CCU.

The sensitivity (ideal value: 1) was >= 0.696 for all the ML models, while it was <= 0.7763 for all the conventional systems. The specificity (ideal value: 1) was >= 0.874 for all the ML models, versus <=

0.868 for all the conventional systems. The PPV (ideal value: 1) for all the conventional systems was <= 0.2244. However, the PPV was > 0.3 for all the ML models. The NPV (ideal value: 1) was very similar among all systems compared. It was >= 0.945 for all the conventional systems, and >= 0.974 for all the ML models. The overall accuracy (ideal value: 1) was different between the conventional systems and the ML models. The maximum for the conventional systems was 0.831 for the LODS. The maximum for the ML models was 0.875 for the XGB and the minimum was 0.866 for the BN (Table 2).

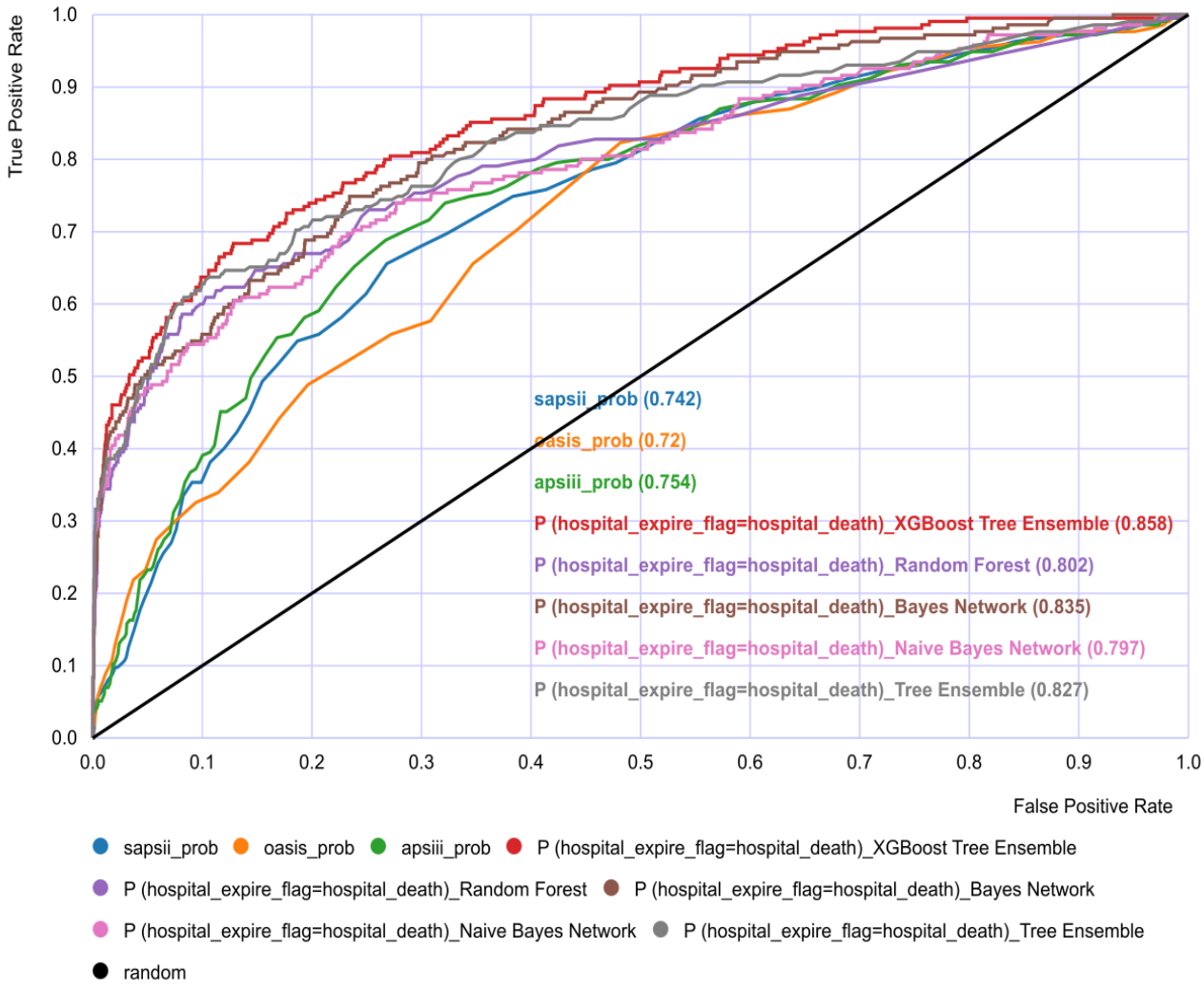


Figure 3. Receiver Operating Characteristic curves for in-hospital mortality prediction for the OASIS, SAPS II, SAPS III severity-of-illness systems, and the Machine Learning models developed. The two-class target variable was *hospital_expire_flag*.

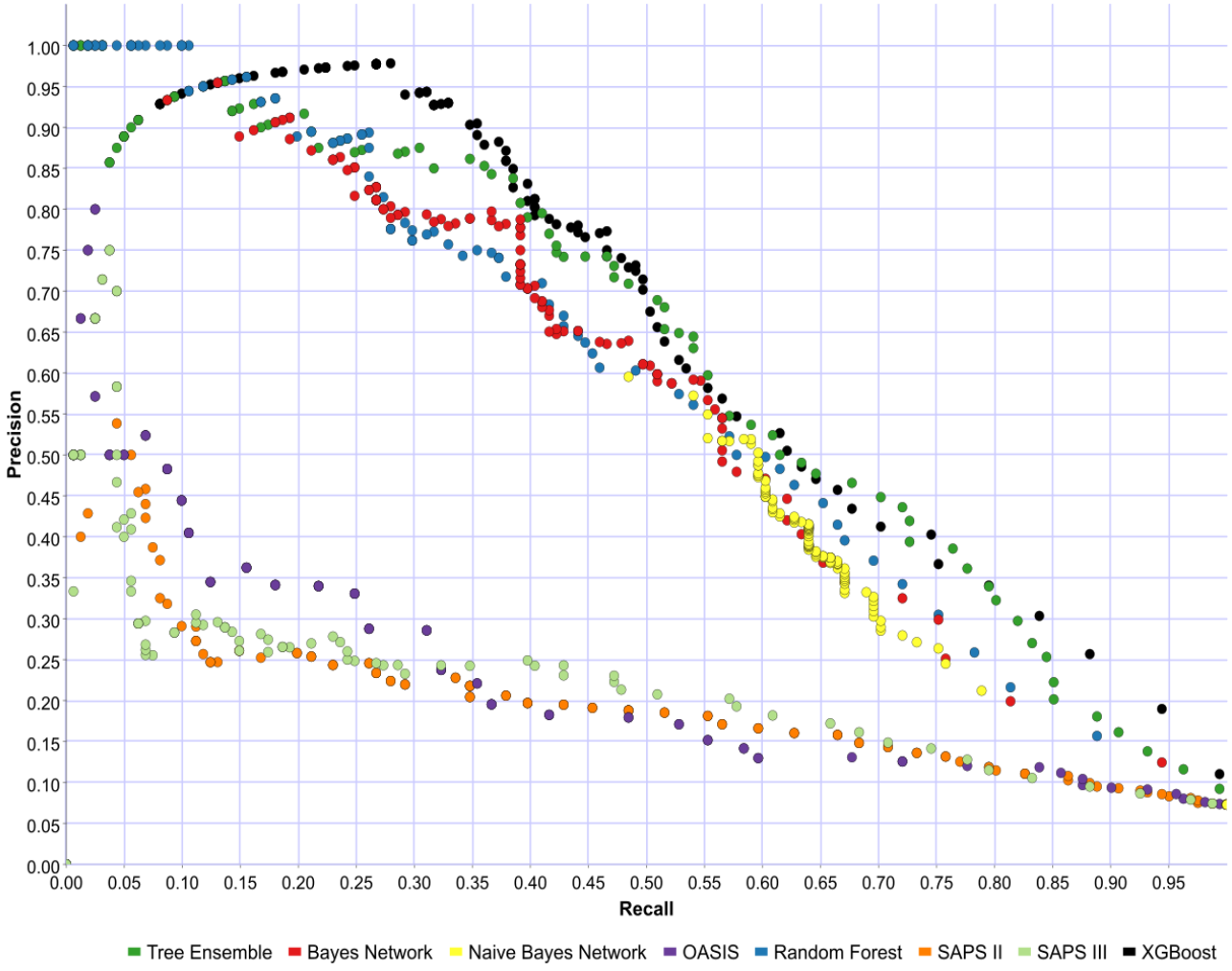


Figure 4. Precision recall curves for mortality prediction in the Cardiac Surgery Recovery Unit and Coronary Care Unit for the SAPS II, OASIS, SAPS III, and the Machine Learning models developed.

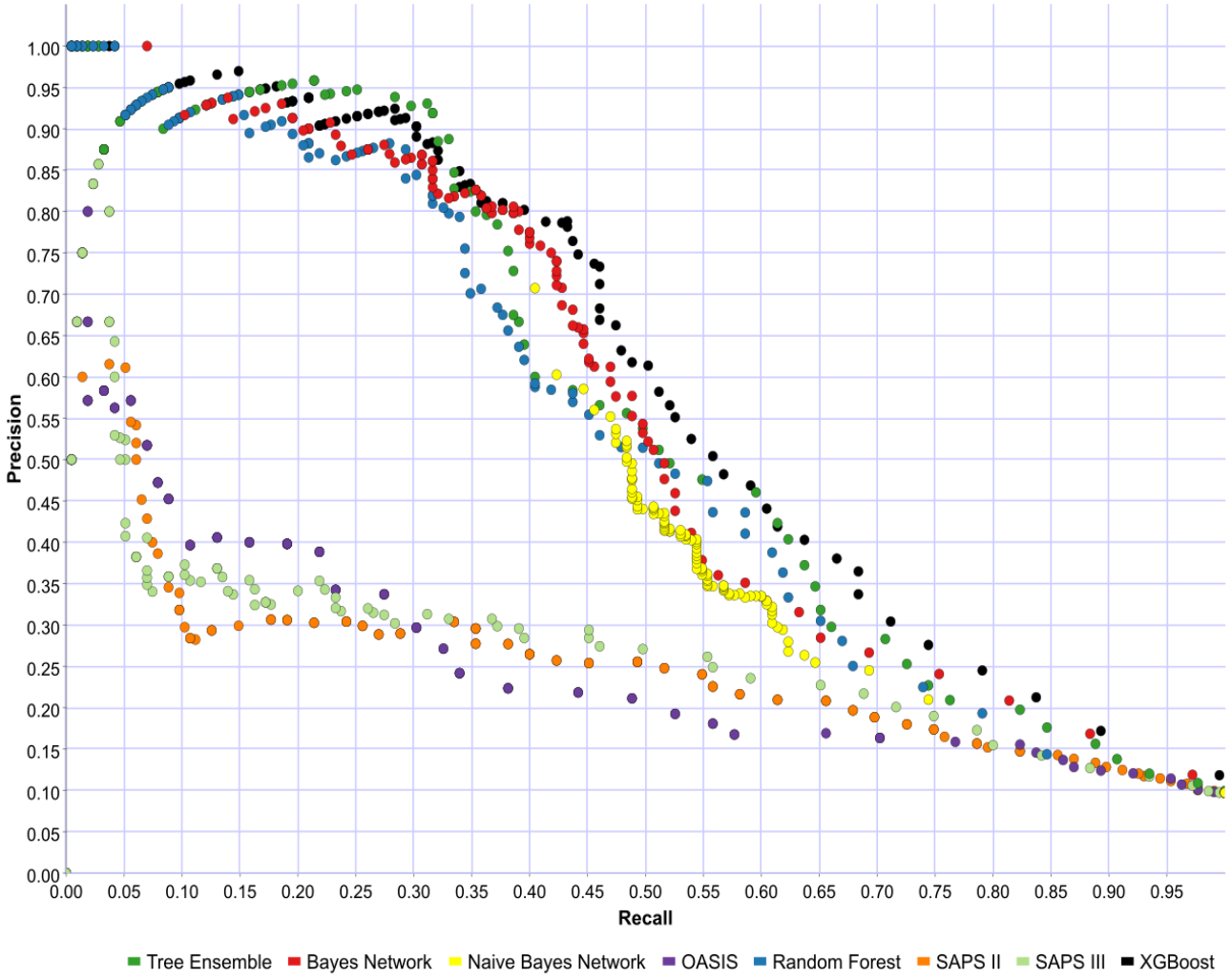


Figure 5. Precision recall curves for in-hospital mortality prediction for the SAPS II, OASIS, SAPS III, Bayes Network, Naive Bayes Network, Tree Ensemble, Random Forest, and XGBoost Tree Ensemble.

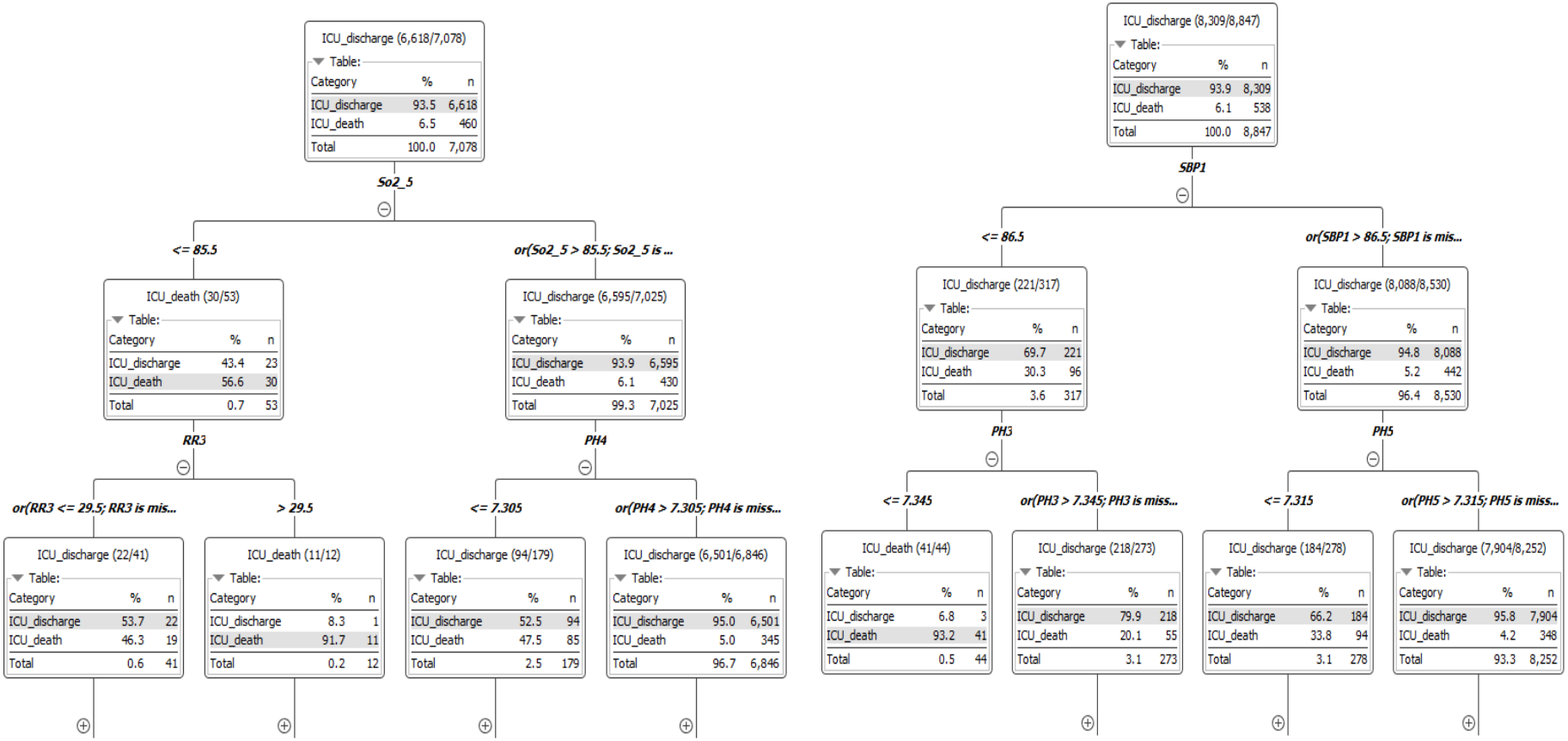
Due to the imbalanced nature of the data, other statistics were evaluated. The DOR (ideal value: infinity) was very different between the conventional systems and the ML models. The DOR was <= 13.012 for all the conventional systems. However, the DOR values were 28.725 for the XGB, 20.323 for the RF, 27.892 for the TE, 18.402 for the BN, and 17.579 for the NB. The CK (ideal value: 1) was quite different between the conventional systems and the ML models. It reached a maximum of 0.2652 for the Serial SOFA, while the maximum for the ML models was 0.422 for the XGB model. The F-measure (ideal value: 1) was also different between the conventional systems and the ML models. It reached a maximum of 0.3481 for the Serial SOFA among the conventional systems, while the maximum for the ML models was 0.48 for the XGB model. Another measure helpful for unbalanced cohorts is BACC (ideal value: 1). The BACC values were more similar to the overall Accuracy for the ML models, while they decreased substantially for conventional systems like the LODS and SAPS III (Table 2).

The MCC values (ideal value: 1) were very different between the conventional systems and the ML

models. The values for the conventional systems ranged between 0.132 and 0.3386, and between 0.414 and 0.47 for the ML models (Table 2). The MK (ideal value: 1) was very different between the conventional systems and the ML models. It ranged between 0.07 for the SAPS and 0.2026 for the Serial SOFA for the conventional systems, and between 0.291 for the BN and 0.326 for the XGB among the ML models (Table 2).

The Brier score (ideal value: 0) for the primary outcome was better for the ML models except the NB, with values between 0.041 for the XGB and 0.084 for the NB. However, it ranged from 0.068 for the SAPS III to 0.12 for the SAPS II among the conventional systems. The values were also better for the secondary outcome for the ML models except the NB, with values between 0.059 for the XGB and 0.107 for the NB. However, it ranged from 0.082 for the SAPS III to 0.124 for the SAPS II among the conventional systems (Table 2).

A DT view for the TE and RF models is provided in Figure 6, showing an exemplar model of each ensemble. In each of these DTs there is a root node, which is a random variable of the features studied. Each node of the tree, except the leaves of the tree at the end (bottom) that represent the values of the utility nodes, represents a random variable of the patient features studied and has several descendants, one for each value of the random variable. In the graphs, the root node is at the top and the descendants are underneath. Each leaf node at the bottom represents the utility conditioned on the values of the nodes above in the corresponding branch. The branches from a random variable node have associated a conditional probability [27,29].



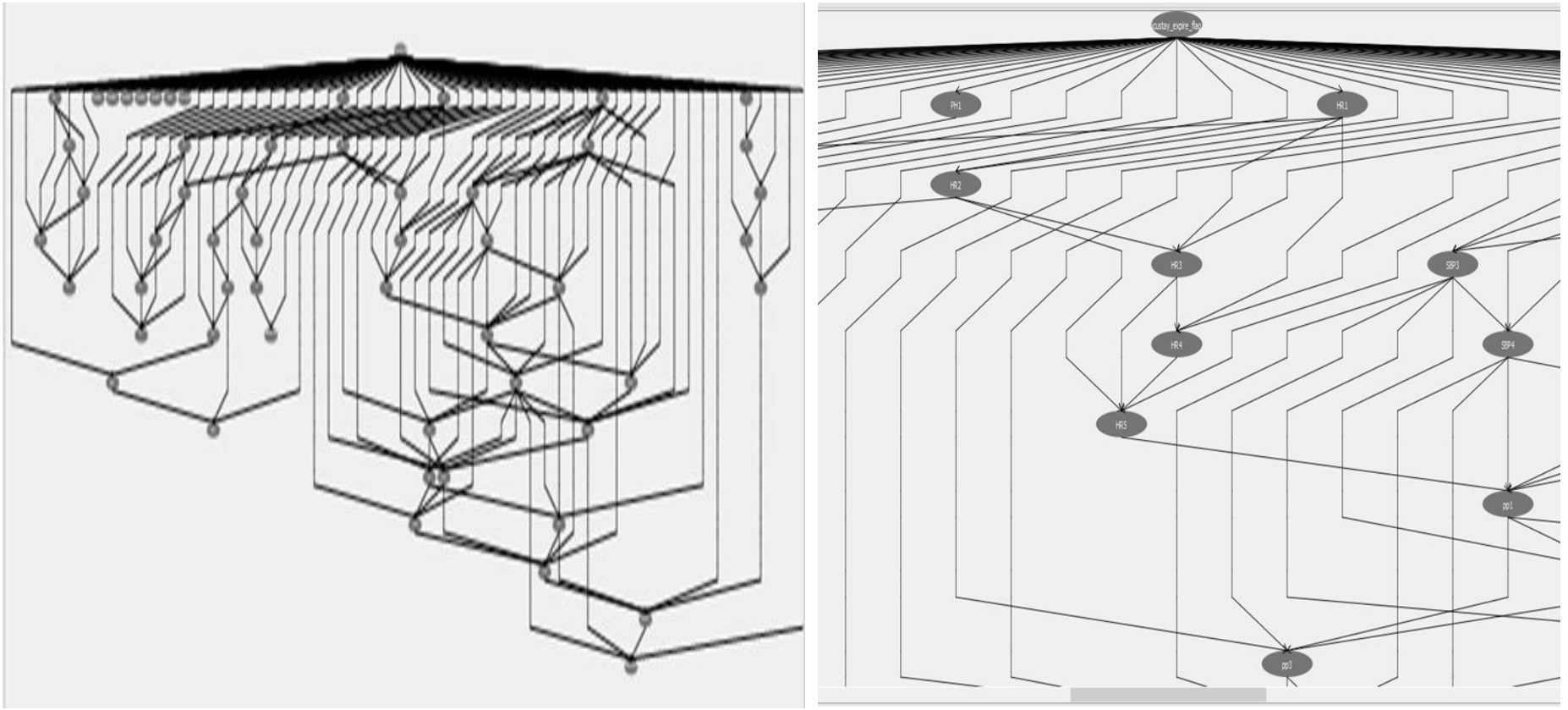
6A **6B**

Figure 6. Decision trees for the primary outcome. In order to closely review the branches of the trees, it is illustrated a shortened version of each tree. **6A:** View of the first decision tree of the 620 trained models in total for the Tree Ensemble of Decision Trees model. **6B:** View of the first decision tree of the 655 trained models in total for the Random Forest of Decision Trees model.

The evaluation of a DT is performed from bottom to top. The utility associated to each branch and each node is calculated taking into account that the utility corresponding to a random variable node is the average of the branches that depart from the node, weighted by the probability. The utility of each final branch is 1 or 0 depending on whether the patient had one value for the feature or the other. The resulting utility of the final root node is the average of the utilities of its two branches, weighted by the probability of each branch [27,29]. This final utility after evaluation of the tree is the probability of mortality for an individual patient.

The DT selected for the TE model of Figure 6A corresponds to the first model of the 620 trained models in total for the primary outcome. The first level decision branch in this tree is the So2 predictor at the 5^{th} hour which separates So2 > 85.5% followed by pH at the 4^{th} hour > 7.305 on the right branch. On the second left branch, an RR at the 3^{rd} hour is split at 29.5 rpm. The DT selected for the RF of Figure 6B corresponds to the first model of the 655 trained models in total for the primary outcome. The first level decision branch in this tree is the SBP predictor at the 1^{st} hour which separates SBP > 86.5 followed by pH at the 5^{th} hour > 7.315 on the right branch. On the second left branch, a pH at the 3^{rd} hour split at 7.345 provides a final leaf at this level, while the other branches continue to branch-out to more levels underneath.

The graph representing the developed BN model for the primary outcome is illustrated in Figure 7, where the parent node represented by the variable *icustay_expire_flag* contains the two output classes and the child nodes are binary nodes associated to the input features.



7A **7B**

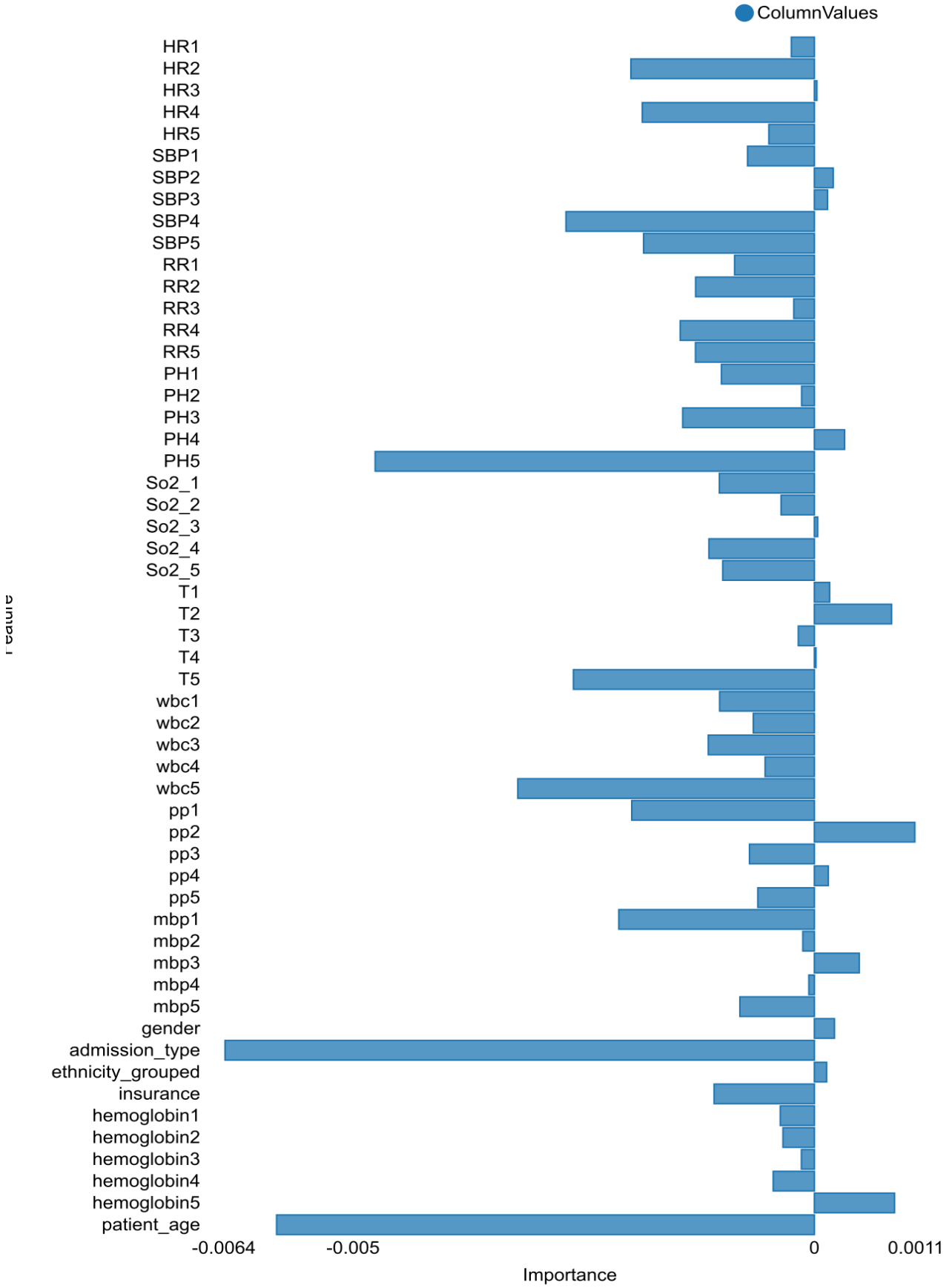
Figure 7. Bayes Network developed for the primary outcome. **7A:** Complete network structure of the Bayes Network classifier. **7B:** An amplified detail of the network showing the target variable *icustay_expire_flag* at the top of the image.

The DTs provided by the ML models may contain some potential to visually depict critical care features and construct a schematic visual logic behind the calculation of the final probabilities of mortality. However, ML models entail more complexity that is not possible to depict graphically completely. Simpler statistical models such as LR provide straightforward models with often heterogeneity in performance, while ML models demonstrate usually higher performance with reduced explainability, such that these ML models are often considered as 'black boxes'.

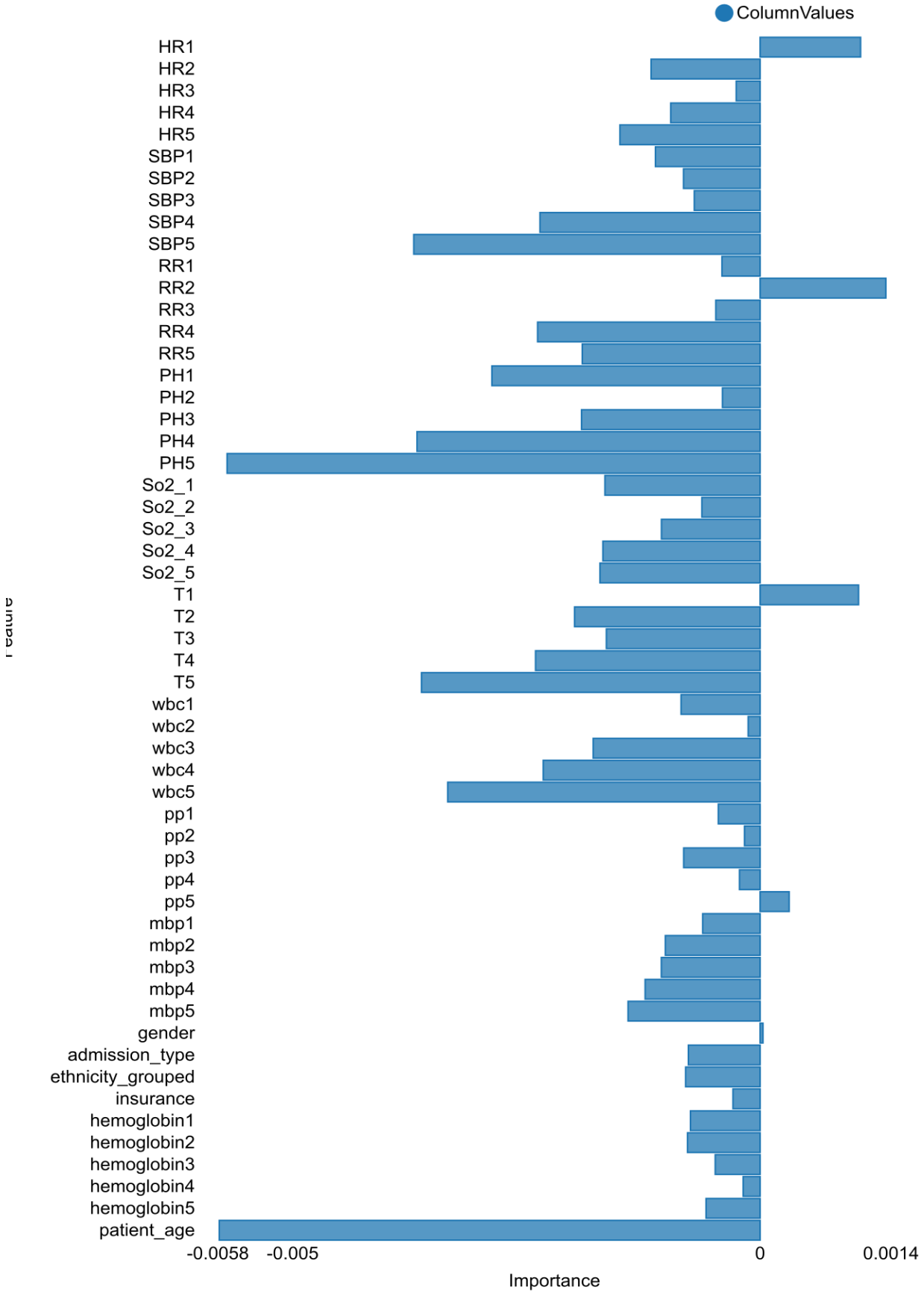
If these ML algorithms are to be implanted for decision-making, then it is necessary for the clinician to understand the logic involved. Algorithms that explain patient-specific predictions have emerged that might increase the understanding of ML prediction models. It was applied in the current developed ML models the algorithm of Shapley additive explanations (SHAP) (see Figure 8).

SHAP is a representation of features´ importance where the contribution of each feature on a prediction is represented using Shapley values. The SHAP value for a feature does not represent a direct and isolated effect of the feature, it is a compound effect as it interacts also with the other features [20]. The algorithm assigns to each feature a Shapley Value that quantifies how much this particular feature changed the output, contributing to the deviation from the mean prediction. The addition of the Shapley values for all features plus the mean prediction equals the actual prediction [26].

The SHAP algorithm was applied to the XGB and TE models to obtain explanations of the features that compose patient-specific predictions for a particular patient (see Figure 8). The ICU record whose prediction of ICU mortality was chosen to be explained corresponded to a male patient of 43.44 years old, admitted as elective to the ICU, who survived to ICU discharge. This patient was correctly predicted by the XGB, which assigned him a probability of discharge of 0.99951. The TE model also correctly predicted the outcome, assigning him a probability of discharge of 0.998387. The Shapley values are depicted in Figure 8 for each feature for the probability of ICU mortality for that patient. It is noteworthy in Figure 8A for the XGB that the young age of 43.44 years old and the admission type as an elective ICU admission were the features with substantially bigger Shapley values, having the greatest contribution towards survival in the context of the other features. This is understandable as young age is usually less associated with mortality and an elective ICU admission is usually less associated with mortality compared to an urgent or emergent ICU admission. As it is observed in Figure 8B for the TE, the young age of 43.44 years old was again a feature with a substantially bigger Shapley value contributing to survival. However, the admission type as elective was captured less strongly as contributor to survival compared to the XGB. There were many similarities with Figure 8A, as the majority of the features pulled towards survival with negative Shapley values.



8A: A correctly classified survivor by the XGBoost Tree Ensemble.



8B: the same correctly classified survivor by the Tree Ensemble.

Figure 8. Algorithm of Shapley additive explanations for the primary outcome for one individual patient. Shapley values are represented on the x-axis showing how much each feature contributed to the probability of ICU mortality for that patient. Features in the bars towards the right of zero in the x-axis favored mortality and those towards the left favored survival. HR=heart rate; SBP=systolic blood pressure; RR=respiratory rate; PH=pH; So2=blood oxygen saturation; T=temperature; wbc=white blood cell count; pp=pulse pressure; mbp=mean blood pressure.

5. DISCUSSION

In this work, ML was applied in the form of NB, BN, TE, RF and XGB models for analysis of patient measurements and demographic features to predict 12-hour advance mortality in the CSRU and CCU, and in-hospital. The ML models developed were able to identify patients at risk for all-factor mortality in the ICU using data extracted from 5 consecutive hours of a patient's ICU stay. The results were compared to six broadly used severity-of-illness scoring systems.

The results revealed small differences in performance for the different ML algorithms analyzed. From the results for prediction of ICU mortality (Table 2), the XGB model showed the highest AUROC followed by the TE, RF, BN, and NB. The conventional systems displayed AUROC values < 0.75 for all the six systems. However, the serial SOFA measured at the same time point as the ML models showed an AUROC of 0.8405. Given that the 24 hours of ICU data used to calculate this serial SOFA are closer to the time of death, an increase in predictive performance was to be expected for serial SOFA. The high values of AUROC for the XGB and TE, together with the nature of their PRC curves indicate that the discriminatory power of these two models for predicting ICU mortality was excellent, substantially outperforming all the conventional systems.

When comparing the AUROC values for the secondary outcome with the AUROC values for the primary outcome, the values for the conventional systems were similar for some and for others they were slightly lower but the values for the ML models were lower for the five models. The order of discriminating ability in descending order was: XGB, BN, TE, RF, and NB. The AUROC value for the serial SOFA was also lower, 0.8093. This more consistent decrease in the AUROC values for the ML models can be explained because the ML models were designed for prediction of the primary outcome. However, some of the conventional systems were originally created for in-hospital mortality prediction instead. These results demonstrate the superiority of the serial SOFA over the static SOFA for the primary and secondary outcomes.

Regarding the PPV, it was low for the ML models but even much lower for the conventional systems. This was due to the very low prevalence of ICU mortality in the cohort. The higher PPV for the ML models can be helpful to decrease the rate of false positives or false alarms, which can decrease alarm fatigue and increase the confidence in a mortality prediction. The NPV was very high for all the conventional systems and ML models, >= 0.945 for all of them. This was influenced by the low prevalence of ICU mortality. The overall accuracy was high for all the ML models, while it was lower for the conventional systems. Due to the unbalanced nature of the cohort, other accuracy measures were evaluated, like the DOR, CK, F-measure, MCC, BACC, and MK.

The DOR was substantially higher for the ML models, in particular for the XGB, TE, and RF. This high DOR for the ML models is very important, as the DOR has been recommended among the best performance measures for unbalanced data-sets [32], as in this work. The BACC was higher for all the

ML models.

The CK was substantially higher for the ML models over all the conventional systems, as well as it was the F-measure and MCC. The MK was also substantially higher for the ML models over the conventional systems (<= 0.2026), particularly for the XGB (0.326), TE (0.315), and RF (0.303). This high MK is important as the literature highlighted this performance metric together with the DOR as the best options to evaluate on unbalanced data-sets, especially in two-class classification problems [32]. However, it was seen in this study that DOR and MK were more sensitive to chosen diagnostic cutoff values than the CK, F-measure and MCC.

The calibration for ICU mortality was better for the ML models over the conventional systems as assessed by the Brier score. However, it was slightly worse for the NB (0.084) than for the OASIS (0.075) and SAPS III (0.068). With regard to the Brier score for in-hospital mortality, the same pattern was observed with better Brier scores for the XGB, TE, BN, and RF over the conventional systems and slightly worse for NB (0.107) compared to OASIS (0.088) and SAPS III (0.082).

Among all the ML models studied, the worse performance with regard to discrimination and calibration was obtained for the NB model. One main reason for this lower performance of NB may be because NB explicitly assumes feature independence, which is violated for the repeated patient measures features collected from the 5 consecutive hours. As the BN was initialized as NB but the maximum number of parents was set as >2 for the BN (43), the resulting model is called a Bayes Net Augmented BN.

Although the accuracy statistics results are dependent on the chosen cutoff values to consider the instance to be classified as positive (ICU death), the fact that the ML models outperformed all the conventional systems across the vast majority of the multiple performance measures studied supports their superiority. In order to reduce false negatives, that is, not to miss patients at high risk of mortality that are classified by the models as having a low risk of mortality, the diagnostic cutoff could be adjusted by decreasing it.

The individual 12-hour mortality prediction ML models developed in this work could be implemented in real-time for a patient, being updated as frequently as needed integrating new observations. This is supported by the use of frequently measured patient clinical variables and routine demographic data.

Limitations of this work include the fact that the data was based on a unique hospital, which may influence the generalization of these models to other cardiac populations. Demographic and institutional differences could affect the performance of the models. Additionally, it must be noted that the used data of the MIMIC-III dates back to between 2001 and 2012. Mortality rates in the CSRU and CCU have reduced over the years since then. The performance of the ML models could be improved in these scenarios by training the models on populations from each medical center before implementation.

While it is understandable that a complex model such as the ML models studied can outperform a simpler one such as the LR used for the conventional systems, improved performance of ML has the drawback of difficulty in explainability. The use of an LR model allows easy explanation of why the risk prediction changed by observing the change in the covariates. However, a similar interpretation cannot easily be produced using ML models, as multiple features will have changed and translating

the impact of these changes within the model is difficult, as well as using complex data transformations that are difficult to interpret.

The explanation algorithm applied in this work provides understanding of how the ML algorithms arrived at the prediction. It is displayed in Figure 8 the contribution of the feature variables to the final patient-specific mortality prediction in a way that is visually explainable. This could facilitate the implantation of ML models into decision-support systems. The DTs provided in Figure 6 may contain some potential also to visually depict interactions among critical care features. Exploring further analysis on individual value ranges selected for each variable to branch-out to the next level could be studied to develop critical care practice guidelines.

The excellent predictive ability of XGB in this work can assist clinicians to anticipate complications and the severity of disease in individual patients, to allocate resources, and to direct attention towards patients at higher risk when under demanding situations at the CSRU and CCU that require prioritization. However, clinical evaluation and medical judgment in the ICU should not be replaced by any predictive tool algorithm. Such algorithms can only support clinicians in decision-making for patient management by providing additional information.

6. CONCLUSIONS AND FUTURE WORK

Overall, for the parameter values provided by the optimization, the developed ML models consistently demonstrated their superiority over the six established systems studied. The discrimination performance for the primary outcome was excellent for XGB, followed by TE, RF, BN, and NB. The calibration was better for the XGB, TE, RF and BN than for the conventional systems. The ML models showed better performance for the vast majority of the accuracy statistics evaluated, most importantly for the measures less sensitive to unbalanced cohorts.

The superiority of the ML models was also appreciated by the fact that the conventional systems require 24 hours of ICU data to generate their scores, except for the SAPS III that requires 1 hour, while the ML models developed require only 5 hours of data to generate their prediction (except labs).

It is shown in this work that traditional severity-of-illness scoring systems can be substantially improved by the use of ML approaches in the setting of the CSRU and CCU, with the XGB model having the best performance among the ones studied.

While an excellent predictive ability was demonstrated for these ML models, in particular for the XGB, it must be emphasized that they are designed as a support tool and they should never be considered in isolation for patient management.

In future work, it could be explored a different design of ML models to exploit a time series analysis led by the dynamic data of ordered time sequences of patient parameter values. This could be accomplished by applying specific types of artificial neural networks that make use of such time sequences like Convolutional neural networks, Recurrent neural networks and Long Short-Term Memory neural networks.

It will be explored also the influence of the underlying (possible) surgeries/procedures performed

prior to and/or during the ICU admission, to study if adding those as features improves the classification.

A version of this study has been accepted and published on 15 April 2021 in the journal named *Journal of Clinical Monitoring and Computing*, with JCR Impact factor 2.108 (2019) (https://doi.org/10.1007/s10877-021-00703-2) [33].

Acknowledgement

A version of this work, authored by Beatriz Nistal Nuño, was first published in *Journal of Clinical Monitoring and Computing* on April 15 of 2021 (doi: 10.1007/s10877-021-00703-2), Epub ahead of print, by Springer Nature. This master's dissertation is reproduced with permission from Springer Nature.

Ethical approval

The data of the MIMIC-III Critical Care Database is of open nature provided access has been granted. The author was formally approved to access the database after completing the required training courses and fulfilling the specific requirements. The project for the MIMIC-III database creation was approved by the Institutional Review Boards of the Beth Israel Deaconess Medical Center in Boston (USA) and the Massachusetts Institute of Technology (USA). Requirement for individual patient consent was waived because the project did not affect clinical care and all protected health information was de-identified.

REFERENCES

1. Zangrillo A, Musu M, Greco T, *et al.* Additive Effect on Survival of Anaesthetic Cardiac Protection and Remote Ischemic Preconditioning in Cardiac Surgery: A Bayesian Network Meta-Analysis of Randomized Trials. *PLoS One*. 2015;10(7):e0134264. doi:10.1371/journal.pone.0134264
2. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, Reinhart CK, Suter PM, Thijs LG. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med*. 1996 Jul;22(7):707-10. doi: 10.1007/BF01709751.
3. Ferreira FL, Bota DP, Bross A, Mélot C, Vincent J. Serial Evaluation of the SOFA Score to Predict Outcome in Critically Ill Patients. *JAMA.* 2001;286(14):1754–1758. doi:10.1001/jama.286.14.1754.
4. Le Gall JR, Loirat P, Alperovitch A, *et al.* A simplified acute physiology score for ICU patients. *Crit Care Med*. 1984;12(11):975-977. doi:10.1097/00003246-198411000-00012
5. Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA.* 1993;270(24):2957–63.
6. Poncet A, Perneger TV, Merlani P, Capuzzo M, Combescure C. Determinants of the calibration of SAPS II and SAPS 3 mortality scores in intensive care: a European multicenter study. *Crit Care.* 2017 Apr 4;21(1):85. doi: 10.1186/s13054-017-1673-6.

7. Haq A, Patil S, Parcells AL, Chamberlain RS. The Simplified Acute Physiology Score III Is Superior to the Simplified Acute Physiology Score II and Acute Physiology and Chronic Health Evaluation II in Predicting Surgical and ICU Mortality in the "Oldest Old". *Curr Gerontol Geriatr Res*. 2014;2014:934852. doi:10.1155/2014/934852

8. Le Gall JR, Klar J, Lemeshow S, Saulnier F, Alberti C, Artigas A, Teres D. The Logistic Organ Dysfunction system. A new way to assess organ dysfunction in the intensive care unit. ICU Scoring Group. *JAMA*. 1996 Sep 11;276(10):802-10. doi: 10.1001/jama.276.10.802.

9. Johnson AE, Kramer AA, Clifford GD. A new severity of illness scale using a subset of Acute Physiology And Chronic Health Evaluation data elements shows comparable predictive accuracy. *Crit Care Med*. 2013 Jul;41(7):1711-8. doi: 10.1097/CCM.0b013e31828a24fe.

10. Isenschmid C, Luescher T, Rasiah R, *et al*. Performance of clinical risk scores to predict mortality and neurological outcome in cardiac arrest patients. *Resuscitation*. 2019;136:21-29. doi:10.1016/j.resuscitation.2018.10.022

11. Adrie C, Cariou A, Mourvillier B, *et al*. Predicting survival with good neurological recovery at hospital admission after successful resuscitation of out-of-hospital cardiac arrest: the OHCA score [published correction appears in *Eur Heart J*. 2007 Mar;28(6):774]. *Eur Heart J*. 2006;27(23):2840-2845. doi:10.1093/eurheartj/ehl335

12. Maupain C, Bougouin W, Lamhaut L, *et al*. The CAHP (Cardiac Arrest Hospital Prognosis) score: a tool for risk stratification after out-of-hospital cardiac arrest. *Eur Heart J*. 2016;37(42):3222-3228. doi:10.1093/eurheartj/ehv556

13. Bisbal M, Jouve E, Papazian L, *et al.* Effectiveness of SAPS III to predict hospital mortality for post-cardiac arrest patients. *Resuscitation*. 2014;85(7):939-944. doi:10.1016/j.resuscitation.2014.03.302

14. Jentzer JC, Bennett C, Wiley BM, *et al*. Predictive Value of the Sequential Organ Failure Assessment Score for Mortality in a Contemporary Cardiac Intensive Care Unit Population. *J Am Heart Assoc*. 2018;7(6):e008169. doi:10.1161/JAHA.117.008169

15. Jentzer JC, Anavekar NS, Bennett C, *et al*. Derivation and Validation of a Novel Cardiac Intensive Care Unit Admission Risk Score for Mortality. *J Am Heart Assoc*. 2019;8(17):e013675. doi:10.1161/JAHA.119.013675

16. Hekmat K, Kroener A, Stuetzer H, *et al*. Daily assessment of organ dysfunction and survival in intensive care unit cardiac surgical patients. *Ann Thorac Surg*. 2005;79(5):1555-1562. doi:10.1016/j.athoracsur.2004.10.017

17. Doerr F, Badreldin AM, Heldwein MB, *et al*. A comparative study of four intensive care outcome prediction models in cardiac surgery patients. *J Cardiothorac Surg*. 2011;6:21. Published 2011 Mar 1. doi:10.1186/1749-8090-6-21.

18. Doerr F, Badreldin AM, Bender EM, *et al*. Outcome prediction in cardiac surgery: the first logistic scoring model for cardiac surgical intensive care patients. *Minerva Anestesiol*. 2012;78(8):879-886.

19. Johnson AEW, Mark RG. Real-time mortality prediction in the Intensive Care Unit. *AMIA Annu Symp Proc*. 2018;2017:994-1003. Published 2018 Apr 16.

20. Thorsen-Meyer HC, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P, Schierbeck J, Strøm T, Chmura PJ, Heimann M, Dybdahl L, Spangsege L, Hulsen P, Belling K, Brunak S, Perner A.

Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *Lancet Digit Health.* 2020 Apr;2(4):e179-e191. doi: 10.1016/S2589-7500(20)30018-2.

21. Jain SS, Sarkar IN, Stey PC, Anand RS, Biron DR, Chen ES. Using Demographic Factors and Comorbidities to Develop a Predictive Model for ICU Mortality in Patients with Acute Exacerbation COPD. *AMIA Annu Symp Proc*. 2018;2018:1319-1328. Published 2018 Dec 5.

22. Nanayakkara S, Fogarty S, Tremeer M, *et al.* Characterising risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study. *PLoS Med*. 2018;15(11):e1002709. Published 2018 Nov 30. doi:10.1371/journal.pmed.1002709.

23. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016 May 24;3:160035. doi: 10.1038/sdata.2016.35.

24. Johnson A, Pollard T, Mark R. MIMIC-III Clinical Database (version 1.4). PhysioNet. 2016. Available from: https://doi.org/10.13026/C2XW26

25. Johnson AE, Stone DJ, Celi LA, Pollard TJ. The MIMIC Code Repository: enabling reproducibility in critical care research. *J Am Med Inform Assoc*. 2018 Jan 1;25(1):32-39. doi: 10.1093/jamia/ocx084.

26. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, *et al.* KNIME: The Konstanz Information Miner. In: Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R, editors. Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization. Berlin: Springer; 2008. p 319-26.

27. Koller D and Friedman N. Probabilistic Graphical Models: Principles and Techniques. The MIT Press, 2009.
28. Spiegelhalter D, Dawid P, Lauritzen S, Cowell R. Bayesian analysis in expert systems. *Stat. Sci.* 8, 219–283 (1993).

29. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. Taylor & Francis, 1984. Taylor & Francis, Jan 1, 1984.

30. Fawcett T. An Introduction to ROC Analysis. Pattern Recognition Letters, Vol. 27, No. 8, 2006, pp. 861-874. doi:10.1016/j.patrec.2005.10.010.

31. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One.* 2015 Mar 4;10(3):e0118432. doi: 10.1371/journal.pone.0118432.

32. Rácz A, Bajusz D, Héberger K. Multi-Level Comparison of Machine Learning Classifiers and Their Performance Metrics. *Molecules*. 2019;24(15):2811.

33. Nistal-Nuño B. Machine learning applied to a Cardiac Surgery Recovery Unit and to a Coronary Care Unit for mortality prediction. *J Clin Monit Comput*. 2021 Apr 15. doi: 10.1007/s10877-021-00703-2. Epub ahead of print.