Artificial Intelligence in Medicine 63 (2015) 191–207

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Predictive modelling of survival and length of stay in critically ill patients using sequential organ failure scores

Rein Houthooft[a,*], Joeri Ruyssinck[a], Joachim van der Herten[a], Sean Stijven[a], Ivo Couckuyt[a], Bram Gadeyne[a,b], Femke Ongenae[a], Kirsten Colpaert[b], Johan Decruyenaere[b,c], Tom Dhaene[a], Filip De Turck[a]

[a] Department of Information Technology (INTEC), Ghent University – iMinds, Gaston Crommenlaan 8, B-9050 Ghent, Belgium
[b] Department of Intensive Care Medicine, Ghent University Hospital, De Pintelaan 185, 2 K12 IC, B-9000 Ghent, Belgium
[c] Department of Internal Medicine, Ghent University, De Pintelaan 185, B-9000 Ghent, Belgium

ARTICLE INFO

Article history:
Received 12 September 2014
Received in revised form 8 December 2014
Accepted 20 December 2014

Keywords:
Mortality prediction
Length of stay modeling
Support vector machines
Critical care
Sequential organ failure score

ABSTRACT

Introduction: The length of stay of critically ill patients in the intensive care unit (ICU) is an indication of patient ICU resource usage and varies considerably. Planning of postoperative ICU admissions is important as ICUs often have no nonoccupied beds available.
Problem statement: Estimation of the ICU bed availability for the next coming days is entirely based on clinical judgement by intensivists and therefore too inaccurate. For this reason, predictive models have much potential for improving planning for ICU patient admission.
Objective: Our goal is to develop and optimize models for patient survival and ICU length of stay (LOS) based on monitored ICU patient data. Furthermore, these models are compared on their use of sequential organ failure (SOFA) scores as well as underlying raw data as input features.
Methodology: Different machine learning techniques are trained, using a 14,480 patient dataset, both on SOFA scores as well as their underlying raw data values from the first five days after admission, in order to predict (i) the patient LOS, and (ii) the patient mortality. Furthermore, to help physicians in assessing the prediction credibility, a probabilistic model is tailored to the output of our best-performing model, assigning a belief to each patient status prediction. A two-by-two grid is built, using the classification outputs of the mortality and prolonged stay predictors to improve the patient LOS regression models.
Results: For predicting patient mortality and a prolonged stay, the best performing model is a support vector machine (SVM) with $G_{A,D} = 65.9\%$ (area under the curve (AUC) of 0.77) and $G_{S,L} = 73.2\%$ (AUC of 0.82). In terms of LOS regression, the best performing model is support vector regression, achieving a mean absolute error of 1.79 days and a median absolute error of 1.22 days for those patients surviving a nonprolonged stay.
Conclusion: Using a classification grid based on the predicted patient mortality and prolonged stay, allows more accurate modeling of the patient LOS. The detailed models allow to support the decisions made by physicians in an ICU setting.

1. Introduction

1.1. Problem statement

The patient length of stay (LOS) is often seen as an indication of the patient resource usage in the intensive care unit (ICU) [1]. Currently ICU physicians generally plan only a single day ahead based on clinical judgement. Automated scheduling assistance based on patient survival and LOS predictions would be beneficial in optimizing ICU resource usage, e.g., estimating the number of occupied beds, as well as individualized patient care. Moreover, this enables the adaptation of surgery scheduling to the predicted ICU load. In addition, predictive ICU models could be a building block in the larger process of making *do not resuscitate* (DNR) decisions to determine whether to stop patient therapy to avoid unnecessary suffering and treatment costs.

In this work, machine learning techniques are trained based on the sequential organ failure (SOFA) score [2–5], a score assessing

* Corresponding author. Tel.: +32 9 331 49 42.
E-mail address: rein.houthooft@intec.ugent.be (R. Houthooft).

http://dx.doi.org/10.1016/j.artmed.2014.12.009
0933-3657/

the daily individual degree of organ failure. The SOFA score is an objective score that allows for calculation of both the number and the severity of organ dysfunction in six organ systems (respiratory, coagulation, liver, cardiovascular, renal, and neurological). The score can measure individual or aggregate organ dysfunction over time and is useful to evaluate morbidity. Although the SOFA scoring was not developed to predict outcome, the obvious relationship between organ dysfunction and mortality has been demonstrated in several studies [3,6,7].

Moreover, patient mortality and LOS estimation is studied in a live monitoring setting by taking into account not only data from the first few days after admission, but also from a moving data window. This allows us to predict the status for a patient with an arbitrary current LOS. Additionally, our models assign a degree of certainty to their classification outputs, allowing ICU physicians to adapt their interpretation of the model to its credibility.

1.2. Related work

In previous studies, ICU patient mortality and LOS modelling has been conducted by taking into account patient data only from day one [8,9]. These studies generally focus on determining whether a patient will have a prolonged stay, i.e., a LOS crossing some predefined threshold. [10] apply machine learning models trained on monitored data from the first five days after patient admission, to predict the patient prolonged LOS, using a 350,000 patient dataset. Contrary to their approach we examine the use of SOFA scores as well as raw data for ICU modelling purposes. SOFA scores are used in a dynamic Bayesian network setting by Sandri et al. [11] to predict sequences of organ failures in a dataset of 79 critically ill patients, however they focus on predicting sequences of organ failures rather than the patient LOS or mortality. Meyfroidt et al. [12] have applied Gaussian processes in ICU patient LOS modelling. They focus on information monitored in the first 4 h after admission and focus on LOS prediction of 960 patients undergoing cardiac surgery. Silva et al. [13] also make use of SOFA scores to build predictive ICU models using a 4425 patient dataset, however their goal is to predict individual organ failures rather than patient mortality, prolonged stay and LOS. Furthermore, [14] have applied a variety of machine learning techniques to model ICU patient survival for a dataset of approximately 1623 patients. However, they focus on a specific patient subset which prevents straightforward generalization of their results.

1.3. Paper organization

The remainder of this paper is structured as follows. In Section 2, we elaborate on the applied predictive models and feature selection methods. Section 3 describes the data used for the applied modelling techniques and sets forth the SOFA score. Hereafter, Section 4 outlines the conducted experiments as well as their results, after which these are discussed in Section 5. Finally, in Section 6 general conclusions are highlighted.

2. Predictive modelling

The survival as well as the prolonged stay prediction are modelled by classification techniques, while the numeric patient LOS is modelled via regression. In this work the following methods are used for classification: artificial neural networks (ANNs) [15], k-nearest neighbors (k-NN) [16], support vector machines (SVMs) [17], classification trees (CART) [18], random forests (RF) [19] and adaptive boosting (AdaBoost) [20]. For regression we use: ANNs, k-NN, RF, support vector regression (SVR) [17], Relevance Vector Regression (RVR) [21] and regression trees (CART) [18]. Some of the experiments are executed using models implemented by SUMO Toolbox [22]. To select the most relevant features both backward elimination and RF, as an importance ranker, are used. In the following paragraphs these applied modelling techniques are described briefly.

2.1. Support vector machines

Support vector machines (SVMs) [17] are sparse kernel machines, a type of models that rely only on a subset of data, the support vectors, to predict unknown values. Additionally, they allow the use of kernels which allow the projection of input data to a different, possibly higher-dimensional space. The model separates the input data by means of a good-fitting hyperplane into two classes. Kernels can be used to transform this hyperplane into a nonlinear input separator, making it a very effective classifier. The SVMs used in this work have the following tunable parameters: a cost term C that controls the misclassification tolerance and acts as a regularization parameter, and one or more kernel parameters.

2.1.1. Probabilistic SVMs

On top of predicting a class, we would like our models to assign a probability, a belief, that a sample is classified correctly. This is done by means of a probabilistic extension of the SVM [23]. As such, a probability

$$P(y = 1|\mathbf{x}) \tag{1}$$

is given for each prediction. This is achieved by – next to optimizing the hyperplane decision boundary – fitting a sigmoid function

$$P(y = 1|\mathbf{x}; A, B) = \frac{1}{1 + \exp\left(Ay(\mathbf{x}) + B\right)} \tag{2}$$

on the decision values y of the SVM classifier. Herein the parameters A and B are estimated by running a maximum likelihood algorithm for Eq. (2) over the original training set.

2.2. Support vector regression

Support vector regression (SVR) [17] is the application of SVMs to regression tasks, in which a linear function is fit through the training set. In this work ϵ-SVR is used, which builds a tube around the fitted curve in which the data points have a zero cost value. Doing so allows us to fit a curve in such a way that many points reside inside this tube. Again, the predictions only depend on a subset of data, the support vectors, which lie on the tube boundaries. Also, kernels can be used to transform the linear fit to a nonlinear curve. The parameters used are the radius ϵ of the tube, which controls the tolerance towards deviation from the fitted curve and acts as a regularization parameter, and one or more kernel parameters.

2.3. Relevance vector machines

SVMs require cross-validation in order to optimally tune their parameters. Furthermore, they cannot capture output uncertainty naturally. Relevance vector machines (RVMs) [21] resemble SVMs, but apply a Bayesian approach to learning by introducing a prior distribution of the SVM weights. They are also sparse as most of the posterior weight distributions concentrate around zero and are hence negligible. The nonzero weights, called relevance vectors, are, unlike SVMs, not based on their distance to a hyperplane or tube. Furthermore, they require less parameter tuning than SVMs, but it can be computationally expensive to train them on large datasets. The regression version of the RVM is called Relevance Vector Regression (RVR).

2.4. Artificial neural networks

Artificial neural networks (ANNs) [15] are machine learning models representable by a graph structure. In this graph the edges represent *weights* and the vertices *neurons*. An ANN consists of multiple layers, next to the input and output layer, called hidden layers which transform data, that is sent through the network from the input features to the output neurons, nonlinearly. In this work, the network is trained by means of back-propagation in which the output is monitored and compared to its correct training value. Hereafter, the error is fed back into the network in order to adjust the weights to obtain a more accurate prediction. The different parameters are the number of layers and the number of neurons in these layers. Furthermore, a weight decay parameter λ is added to restrict the model complexity. This parameter makes sure that weights having have no substantial effect on the predictive power converge to zero to avoid overfitting.

2.5. k-Nearest neighbours

k-Nearest neighbors (k-NN) is a learning algorithm capable of classification as well as regression. In k-NN the output is determined by the label of the k closest data points in the feature space. It can be seen as a lazy learning algorithm as the k closest points (in our case based on the Euclidean distance) are calculated at query time. Hence the model can easily be used in an online learning setting. In k-NN classification the mode of the class of the k nearest neighbours is used, while in the regression case the model uses the average neighbour value.

2.6. Classification and regression trees

Decision trees are eager learning algorithms that generalize training data by building a tree structure. In this tree every node represents a split on a certain property of a data sample. By following a path corresponding to the to-be classified sample from the root to a leaf node along the edges corresponding to the properties of this sample, it is possible to predict property values. Classification And Regression Trees (CART) [18] is such a decision tree algorithm capable of classification as well as regression. It is binary in the sense that every node has exactly two child nodes. Furthermore, multiple nodes along a root-leaf path may split on the same property. At each node of the tree, the node split property is chosen based on the maximal decrease in impurity, an information entropy measurement.

2.7. Random forests

Random forests [19] is an ensemble machine learning method based on the construction of multiple CART decision trees for either regression or classification. The main underlying technique used in random forests is *bootstap aggregating* (bagging). In bagging the training data set is sampled K times and each time S samples are taken with replacement. These K different subsets are then used to train K different CART decision trees. Furthermore, each of the K CART learners only use $|\mathbf{x}|^{1/2}$ features for a dataset of $|\mathbf{x}|$ features in order to reduce the correlation of strong predictors. The predicted value of the random forests algorithm is the mode in case of classification or the average value of the K different decision trees in case of regression.

2.8. Discrete adaptive boosting

Discrete adaptive boosting (Discrete AdaBoost) [20] is a boosting algorithm that linearly combines multiple weak learners to form a single strong one applicable to binary classification problems. The weak learners may be chosen arbitrarily, but the ability to train them in a weighted manner (e.g., an SVM with class weights) is required. The different weak learners are trained sequentially in such a way that each subsequent weak learner tries to correctly classify the samples wrongly predicted by the preceding weak learners by increasing their weights. In this work we use CART as a weak classification algorithm.

2.9. Feature selection

The feature importance ranking is obtained by running a variant of the RF [19] algorithm, tailored to obtaining feature importance. For each tree l_k generated by the RF algorithm, about 1/3th of the data is left out (see Section 2.7), which is called the out-of-bag (OOB) data. After building l_k based on the non-OOB data, the OOB data is used as a test set. As such, each data sample is used as a test sample in $\pm$1/3th of the total number of trees K. Hereafter, the number of times a tree l_k makes incorrect predictions is divided by K. This number is then averaged over all data samples and is called the OOB error. Next, for each tree l_k as well as for each feature, the feature values in the OOB data are randomly permuted after which the previous error estimating algorithm is run again. The difference between the two OOB error rates is scaled by the standard deviation of the differences, and is called the *importance* of a feature.

An advantage of this algorithm is its ability to visualize the importance of each distinct features, a downside, however, is its inability to detect correlations between features. To overcome this, features can be removed by means of *backward elimination*. In backward elimination the least informative features are sequentially removed. As such, it is possible to obtain a minimal set of features that is still able to get maximal classification or regression performance.

Furthermore, we use Sobol indices [24] to perform a sensitivity analysis on our models. These indices are a Monte Carlo sensitivity analysis [25] which explain the importance of an input feature based on the variance of the output. In particular, it shows the the variance of the conditional expectation of the output, given a particular input feature value, normalized by the total variance of the output. As such, the result is a percentage, indicating how much of the model variance is explained by a particular feature. Higher values correspond to features of higher importance.

3. Collected dataset

First, we describe the data used for training the machine learning models. Next, we explore the details of the SOFA score calculations, the use of raw data values as training features, as well as the use of a moving data window.

3.1. Data collection

Our dataset consists of all patient admitted between January 1, 2009 and September 17, 2013 to the 22 bed surgical ICU and 14 bed medical ICU of the Ghent University Hospital. This data contains individual patient information, ICU stay information and ICU monitored values, as well as lab test results. In total 18,921 patients were recorded in the dataset. The patient information consists of five attributes: the ICU admission time, ICU discharge time, weight, age and sex. Furthermore, all parameters required to calculate the SOFA scores are measured. These consist of 13 status parameters and 6 parameters describing the administered drugs, all accompanied by a timestamp.

The data is pre-processed by removing all underage (<18 years) patients, retaining 14,480 patients of which 38.1% is female and 61.9% is male. From this total set of patients, 91.8% survive their ICU stay. Parameters are not monitored when ICU physicians determine

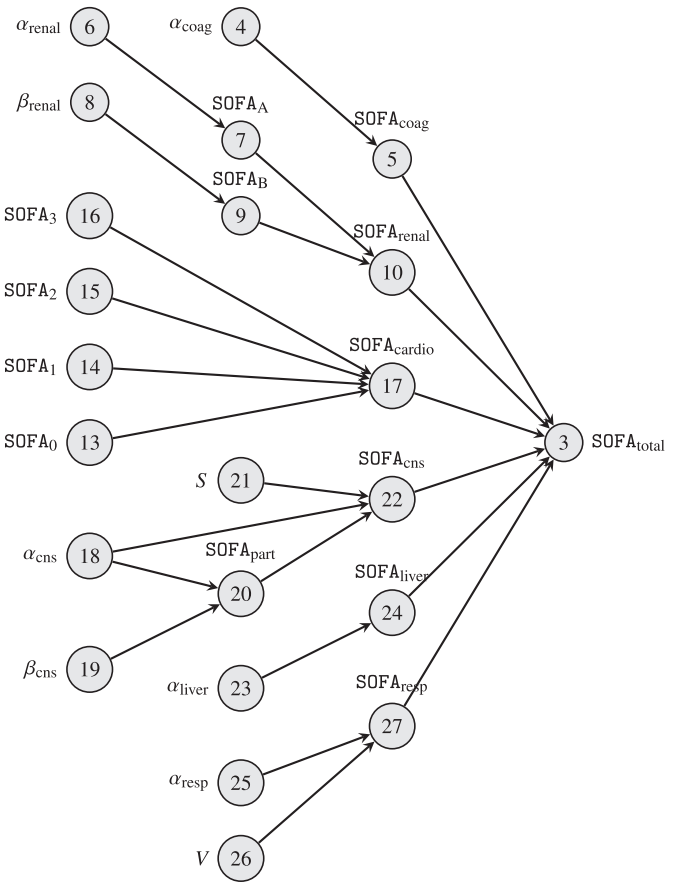


Fig. 1. Graphical summary SOFA score calculation: each node denotes a specific calculation, the number inside each node corresponds to the equation number which explains how the computation is done; the arrow heads represent calculation inputs for each node, while the opposite ends represent the calculation output for each node.

that a particular patient organ status is not affected at all. However, the SOFA computation methods make up for this lack of data.

3.2. SOFA score calculation

The data parameters, for which the sampling frequency is variable as the number of measurements per day for a particular parameter is not fixed, are aggregated into a single daily representative score, called the sequential organ failure (SOFA) score [2–5]. This score quantifies the patient's organ performance and is generally used to assess the patient's status during his ICU stay. This SOFA score is based on six SOFA sub-scores calculated for: the coagulation function, renal function, cardiovascular system function, central nervous system function, hepatic system function and respiratory function. These sub-scores are integers within the range [0, 4]. A low SOFA score indicates a healthy patient status, while a high score may indicate organ failure. For each patient a SOFA score is calculated for each organ system function, every day after admission, at 5:00 AM. This score aggregates different data values measured within the 24 h window, denoted as W, preceding each daily SOFA score calculation. As such, W =](5 : 00 AM day$i-1$), ..., (5 : 00 AM day i)]. When doctors do not suspect a patient having a certain critical organ status, no values are measured for that particular organ system (e.g., no liver test is conducted). To deal with these missing data, the corresponding SOFA sub-score is set to zero, which denotes a healthy status. The calculation method of the different SOFA sub-scores is explained in the following sections. A graphical summary of the SOFA score calculation is given in Fig. 1.

3.2.1. SOFA score

The total SOFA score is be calculated by adding six SOFA sub-scores of each different organ system, which will be explained in the next sections:

$$SOFA_{\text{total}} = SOFA_{\text{coag}} + SOFA_{\text{renal}} + SOFA_{\text{resp}} + SOFA_{\text{cardio}} + SOFA_{\text{cns}} + SOFA_{\text{liver}}. \quad (3)$$

This total SOFA score is generally used by physicians to assess the patient status. Its range is $0 \leq \texttt{SOFA}_{\text{total}} \leq 24$ as every sub-score has a range of [0, 4]. Later in this text, a SOFA score accompanied by a number in subscript denotes the day after admission it corresponds to. For example $\texttt{SOFA}_{\text{renal(3)}}$ represents the renal system function SOFA score on day 3 after admission.

3.2.2. SOFA subscore 1: coagulation function

The coagulation function SOFA score is calculated by measuring the trombocytes [number of platelets $\times 10^3/\mu$l] value in the blood,

$$\alpha_{\text{coag}} = \min_{t \in W}\{trombocytes(t)\}. \quad (4)$$

This minimum trombocytes value α_{coag} of all measurements at a time $t \in W$ is used in

$$SOFA_{\text{coag}} = \begin{cases} 4 & : \alpha_{\text{coag}} \leq 20 \\ 3 & : \alpha_{\text{coag}} \leq 50 \\ 2 & : \alpha_{\text{coag}} \leq 100 \\ 1 & : \alpha_{\text{coag}} \leq 150 \\ 0 & : otherwise \end{cases}, \quad (5)$$

which defines the coagulation function SOFA score.

3.2.3. SOFA subscore 2: renal function

The renal function SOFA score is based on two input values: the plasma creatine [mg/dl] value and the urine volume [ml]. To make sure the urine volume value is a correct reflection of the 24 h urinary output, it has to be guaranteed that the patient was not admitted within W. First, α_{renal} is calculated as

$$\alpha_{\text{renal}} = \max_{t \in W}\{plasma\,creatine(t)\}, \quad (6)$$

which is used in

$$SOFA_{\text{A}} = \begin{cases} 0 & : \alpha_{\text{renal}} < 1.2 \\ 1 & : \alpha_{\text{renal}} < 1.9 \\ 2 & : \alpha_{\text{renal}} < 3.4 \\ 3 & : \alpha_{\text{renal}} < 4.9 \\ 4 & : otherwise \end{cases} \quad (7)$$

to calculate a first partial SOFA score. Next, the total urine volume β_{renal} is calculated as

$$\beta_{\text{renal}} = \sum_{t \in W} urine(t), \quad (8)$$

which is used in

$$SOFA_{\text{B}} = \begin{cases} 3 & : \beta_{\text{renal}} < 500 \wedge T_{\text{admission}} \notin W \\ 4 & : \beta_{\text{renal}} < 200 \wedge T_{\text{admission}} \notin W \\ 0 & : otherwise \end{cases} \quad (9)$$

to calculate a second partial SOFA score. The renal SOFA score is then defined as the maximum of these two partial scores,

$$SOFA_{\text{renal}} = \max\{SOFA_{\text{A}}, SOFA_{\text{B}}\}. \tag{10}$$

3.2.4. SOFA subscore 3: cardiovascular system function

To calculate the cardiovascular system function SOFA score, first the mean arterial pressure (MAP) is calculated by the invasive mean arterial blood pressure (IBP) [mm Hg]:

$$MAP = \min_{t \in W}\{IBP(t)\}. \tag{11}$$

If no IBP value exists within the window W, the noninvasive mean arterial blood pressure (NIBP) value [mm Hg] is used:

$$MAP = \min_{t \in W}\{NIBP(t)\}. \tag{12}$$

This MAP value is used in a first partial SOFA score:

$$SOFA_0 = \begin{cases} 0 & : MAP > 70 \\ 1 & : otherwise \end{cases} \tag{13}$$

The maximum administered dopamine (DOP) value [μg/kg/min] is used in a second partial SOFA score:

$$SOFA_1 = \begin{cases} 2 & : 0 < \max\limits_{t \in W}\{DOP(t)\} \leq 5 \\ 3 & : 5 < \max\limits_{t \in W}\{DOP(t)\} \leq 15 \\ 4 & : otherwise \end{cases}, \tag{14}$$

and the administered dobutamine (DOBU) value [μg/kg/min] is used in a third one:

$$SOFA_2 = \begin{cases} 2 & : \max\limits_{t \in W}\{DOBU(t)\} > 0 \\ 0 & : otherwise \end{cases} \tag{15}$$

Both the maximum administered epinephrine (EPI) [μg/kg/min] and the maximum administered norepinephrine (NOREPI) [μg/kg/min] are used in a fourth partial SOFA score:

$$SOFA_3 = \begin{cases} 3 & : (\max\limits_{t \in W}\{EPI(t)\} \leq 0.1) \vee (\max\limits_{t \in W}\{NOREPI(t)\} \leq 0.1) \\ 4 & : (\max\limits_{t \in W}\{EPI(t)\} > 0.1) \vee (\max\limits_{t \in W}\{NOREPI(t)\} > 0.1) \end{cases} \tag{16}$$

The total cardio SOFA score is then computed as the maximum of the four partial SOFA scores:

$$SOFA_{cardio} = \max_{i \in \{0,1,2,3\}}\{SOFA_i\}. \tag{17}$$

Note that the cardio SOFA score calculation is more complex than the others. The blood pressure is used to give a first indication in case no drugs are administered. When drugs are administered, they override the blood pressure-based partial SOFA score, causing the SOFA score to be adjusted based on the amount of drugs administered.

3.2.5. SOFA subscore 4: central nervous system function SOFA score

For calculating the SOFA score of the nervous system function, two coma Glasgow score (CGS) values are used, namely a derived and a monitored value. When the patient is not sedated, the derived value is used, defined as

$$\alpha_{\text{cns}} = \min_{t \in W}\{CGS(t)\}. \tag{18}$$

In the other case, the monitored value is used, regardless of the time window W. This monitored value is the last known derived CGS value before sedation:

$$\beta_{\text{cns}} = CGS_{\text{monitored}}(t_{\text{last}}). \tag{19}$$

A partial SOFA score is defined as

$$SOFA_{part}(x) = \begin{cases} 4 & : x < 6 \\ 3 & : x \leq 9 \\ 2 & : x \leq 12 \\ 1 & : x \leq 14 \\ 0 & : otherwise \end{cases}. \tag{20}$$

Furthermore, we check whether the patient was sedated (S) within the time period W. This is done by examining whether any sedatives were administered:

$$S = \exists t \in W : \begin{array}{l} (Diprivan\,1\%(t) > 0) \vee (Diprivan\,2\%(t) > 0) \\ \vee (Dormicum\,15mg(t) > 0) \\ \vee (Dormicum\,50mg(t) > 0) \end{array}. \tag{21}$$

The central nervous system SOFA score can then be calculated as

$$SOFA_{\text{cns}} = \begin{cases} SOFA_{\text{part}}(\alpha_{\text{cns}}) & : \alpha_{\text{cns}} \leq 12 \wedge S \\ SOFA_{\text{part}}(\beta_{\text{cns}}) & : otherwise \end{cases}. \tag{22}$$

3.2.6. SOFA subscore 5: hepatic system function

The hepatic system function SOFA score is calculated by measuring the maximum bilirubine serum value [mg/dl] within the 24 h window W:

$$\alpha_{\text{liver}} = \max_{t \in W}\{bilirubine(t)\}. \tag{23}$$

This maximum value is then used to calculate the hepatic system function SOFA score,

$$SOFA_{\text{liver}} = \begin{cases} 4 & : \alpha_{\text{liver}} \geq 12 \\ 3 & : \alpha_{\text{liver}} \geq 6 \\ 2 & : \alpha_{\text{liver}} \geq 2 \\ 1 & : \alpha_{\text{liver}} \geq 1.2 \\ 0 & : otherwise \end{cases} \tag{24}$$

3.2.7. SOFA subscore 6: respiratory function SOFA score

The respiratory function SOFA score is calculated by measuring the minimum PaO_2/FiO_2-ratio (PF) [mm Hg] within the 24 h window W:

$$\alpha_{\text{resp}} = \min_{t \in W}\{PF(t)\}. \tag{25}$$

Furthermore, a parameter V, describing whether the patient was ventilated or not, is calculated as

$$V = \exists t \in W : \begin{array}{l} (RRv_s(t) > 0) \vee (RRv_m(t) > 0) \vee (VE_s(t) > 0) \\ \vee (VE_m(t) > 0) \vee (PEEP_s(t) > 0) \vee (PEEP_m(t) > 0) \end{array}. \tag{26}$$

Herein is RRv the respiratory rate ventilator frequency, PEEP the positive end expiratory pressure, the pressure added during exhalation, and VE the expiratory volume. The subscript s denotes the input settings on the ventilation device while the subscript m denotes the measured values. These two values are then consolidated into the respiratory function SOFA score,

$$SOFA_{\text{resp}} = \begin{cases} 4 & : \alpha_{\text{resp}} \leq 100 \wedge V \\ 3 & : \alpha_{\text{resp}} \leq 200 \wedge V \\ 2 & : \alpha_{\text{resp}} \leq 300 \\ 1 & : \alpha_{\text{resp}} \leq 400 \\ 0 & : otherwise \end{cases}. \tag{27}$$

3.3. Raw data values

To evaluate the added value of model training via SOFA scores, models are also trained on the underlying raw data. These raw data values are: α_{coag} for the coagulation system, α_{renal} and β_{renal} for the renal system, α_{cns} and S for the central nervous system, α_{liver} for the hepatic system, and α_{resp} and V for the respiratory system. Only the cardiovascular SOFA score was not split into underlying values due to its complex aggregating function. Additionally, the daily fluid balance (`fluid`) [ml] was as a raw data value, which is not present in any of the SOFA sub-scores. In case of missing parameters, these raw values were set to values corresponding to a healthy patient status. This can be motivated by the fact that physicians do not measure a specific organ value if they do not suspect a critical status for that particular organ. As with the SOFA score representation, a raw value accompanied with a subscript represent the day after admission it belongs to.

3.4. Moving window

A secondary dataset was created by artificially extending the first one via a moving window. This dataset consists of data from the days $\{(1+i),\ldots,(5+i)\}$ with $i \in \{0,\ldots,(\text{LOS}-5)\}$. Doing so increases the number of data samples from 3288 to 33,630, allowing the simulation of live predictions with respect to the patient mortality, prolonged stay and numerical LOS. As such, we can model a patient with an arbitrary current LOS, enabling daily predictions by observing a patient during his stay. Moreover, this allows the models to be trained on this enriched data in order to predict a patient's future status, 5 days after admission, more accurately.

4. Methodology and results

In this section, the different models are evaluated according to their predictive power regarding patient survival, prolonged stay, and remaining LOS, denoted as Φ. We also check whether using the underlying raw values rather than using SOFA scores improves the prediction accuracy. Additionally, the models are tested on retrospective data as if it were live patient data, taking into account patient data not only from the first five days, but from the previous five to make moving window predictions.

4.1. Methodology

To measure the performance of the different classification models, the *recall* and *precision* of both classes (in our case classification is always binary), C_1 and C_2, are measured. The recall r_{C_k} is the ratio of samples in the validation set having a class C_k that are also identified by the classifier, for each distinct class. The precision p_{C_k} is measured as the the ratio of correctly predicted classes, for each distinct class. Note that the definition of the p_{C_1} is equivalent to the definition of the *specificity* of class C_2, while r_{C_1} is equivalent to the *sensitivity* of class C_1. In order to measure the effectiveness of the regression techniques modelling Φ, the average offset between the predicted number of days and the correct number of days, denoted as $\bar{\Delta}\Phi$, and the median, denoted as $\hat{\Delta}\Phi$, are evaluated. However, due their nature, simply averaging all recall and precision measurements is not an option (e.g., obtaining a 100% value for recall (or precision) and a 0% value for precision (or recall), leading to an average of 50%, should be avoided). To strike a balance between the different measurements, the models are optimized for the geometric mean of all quantities, defined as

$$G_{C_1,C_2} = (r_{C_1} r_{C_2} p_{C_1} p_{C_2})^{1/4}. \tag{28}$$

For training the different machine learning models, *repeated random sub-sampling validation* (RRSSV) [26], in which the dataset is split n times in a training set (60%) and a validation set (40%), is used. Over these n splits, the average or median of the measured values (e.g., median offset or average recall) is computed. Model parameters are sought by a grid search, defined by a lower bound L, an upper bound U and a step size s. Each parameter to be tested is drawn from the set $\{L, L+s, L+2s, \ldots, U\}$. In case of multiple parameters, all their combinations are explored. The following parameter spaces were searched:

- RF: For the number of CART trees: $L=10$, $U=1000$ and $s=10$.
- SVM: With C as a function of 10^n, for n: $L=-4$, $U=4$ and $s=1$; with γ as a function of 10^n, for n: $L=-4$, $U=4$ and $s=1$.
- ANN: 3-layer network (number of layers remained fixed), for the number of hidden neurons: $L=1$, $U=50$ and $s=1$; for the decay rate λ: $L=0.1$, $U=0.9$ and $s=0.2$.
- k-NN: For the number of neighbors k: $L=10$, $U=500$ and $s=10$.
- ADA: For number of modified CART trees: $L=10$, $U=200$ and $s=10$.
- SVR: With C as a function of 10^n, for n; $L=-4$, $U=4$ and $s=1$; with γ as a function of 10^n, for n: $L=-4$, $U=4$ and $s=1$; with ϵ as a function of 10^n, for n: $L=-10$, $U=-4$ and $s=1$.
- RVR: With γ as a function of 10^n, for n: $L=-4$, $U=4$ and $s=1$.

After the model is tuned using this approach, the final results are obtained in an RRSSV setting as averages over all validation iterations. These results are also compared to a baseline approach (denoted as base in the result tables). This baseline predicts outputs by randomly sampling the validation or test set.

The data used for the modelling experiments are based on the fraction of patients that were still present in the ICU after the calculation of the SOFA score on day 5 after admission. As such, 3287 patient observations remain, which comprise the full dataset for training and validation purposes. Furthermore, all raw feature data are normalized to the interval [0, 1], while SOFA score features keep their [0, 4] range.

A paired Wilcoxon signed-rank test is performed on the results, comparing their statistical difference in a pairwise manner. The significance level was set to $p=0.05$. In case two results are not statistically significant, this is mentioned with a superscript symbol in the corresponding result table.

4.2. Patient mortality prediction

Fig. 2 depicts the relation between the (total) SOFA score on day 5 and the survival rate of a patient. The figure shows that an increasing total SOFA score corresponds to decreasing patient survival chances. This might indicate the predictive power of SOFA scores regarding patient mortality. Moreover, this figure shows the empirical cumulative density function (ECDF) of this score. The SOFA scores between 0 (the minimum) and 12 are approximately uniformly distributed, pointed out by the constant curve slope, while only a minority of the total patient set has a SOFA score between 12 and 24 (the maximum).

Because the data is distributed unevenly in terms of patient mortality, e.g., only a small fraction of the patients do not survive their ICU stay (8.2%), re-sampling with replacement is used to redistribute the training samples of the different classes. This eliminates the bias towards predicting the most represented class correctly. The validation set on the other hand remains unaltered.

Before evaluating the models (SVM, ANN, RF, k-NN and AdaBoost), their parameters are optimized regarding $G_{A,D}$ (see Eq. (28)), while restricting the model complexity, by running a coarse-grained grid-search with RRSSV ($n=50$) over the parameter space,

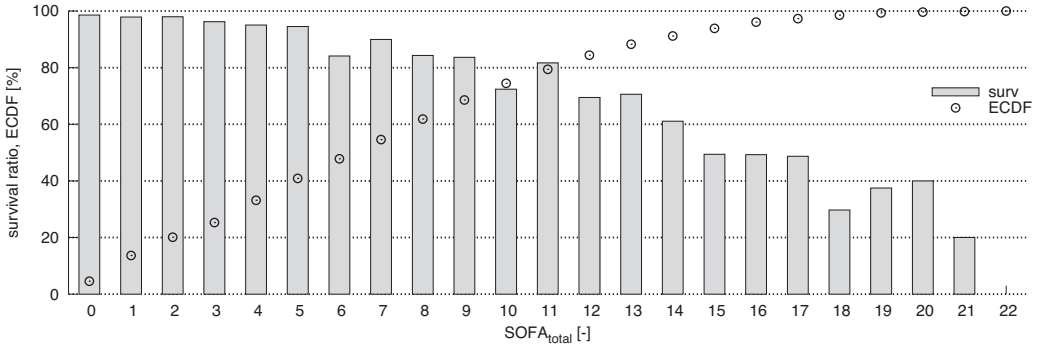


Fig. 2. Correlation between $\mathtt{SOFA}_{\mathtt{total(5)}}$, the SOFA score on day 5, and the patient survival rate. This rate is shown together with the ECDF of the number of patients for each $\mathtt{SOFA}_{\mathtt{total(5)}}$ value.

as explained in Section 4.1. This leads to the following parameter values for each model:

- RF: 150 CART classification trees ($G_{A,D}$ converges above 150 trees).
- SVM: $C = 10$, Gaussian radial kernel with $\gamma = 0.001$; probabilistic outputs generated by a fitted sigmoid function.
- ANN: 3-layer (number of layers was fixed) network with 20 hidden neurons and 1 output neuron; weight decay set to $\lambda = 0.5$.
- k-NN: regression via the average of the $k = 150$ nearest neighbours, based on the Euclidean distance.
- ADA: 50 modified CART trees are used as weak classifiers.

Before gathering results, a pre-run searching for the best performing model is executed on all patient data (both SOFA scores and raw values). Next, the most important/predictive features are sought by running an RF importance ranking algorithm (see Section 2.9). Hereafter, the added value of the most important features is investigated by sequentially removing them via backward elimination (see Section 2.9). This latter feature selection stage is conducted using the best-performing classifier selected in the pre-run, namely the SVM.

The results of the RF feature importance ranking algorithm are shown in Fig. 3 for the different SOFA scores as well as the raw data values. In this plot the values are grouped in groups of 5 days (left to right: day 1–5). Higher scores indicate more important features as randomly permuting their values has a larger influence on the classification accuracy. However this importance estimate is naive as it does not take into account correlations between features, it only shows how much adding irrelevant data deteriorates the classifier. Nonetheless, a general trend can be witnessed in which features of days which are closer to the current day are of greater importance.

To overcome this downside of feature correlation, each feature is sequentially removed from the classifier via backward elimination. The most important features are removed first while monitoring the change in $G_{A,D}$. Doing so allows us to identify a minimum set of features capable of predicting the patient survival rate with maximum accuracy. The results from this experiment are shown in Fig. 4. In terms of SOFA score features, $\mathtt{SOFA}_{\mathtt{resp(4)}}$, $\mathtt{SOFA}_{\mathtt{cardio(5)}}$, $\mathtt{SOFA}_{\mathtt{renal(5)}}$, $\mathtt{SOFA}_{\mathtt{coag(2)}}$, $\mathtt{age}$ and $\mathtt{SOFA}_{\mathtt{cns(5)}}$ are sufficient to grant the model its maximum accuracy. In terms of raw data, a larger set of different features contributes to the model's predictive power. Herein, $\mathtt{cns}_{2,4,5}$, $\mathtt{PF}_5$, $\mathtt{SOFA}_{\mathtt{cardio(1)}}$, $\mathtt{bilirubine}_5$, $\mathtt{ventilated}_2$, $\mathtt{urine}_{4,5}$, $\mathtt{age}$ and $\mathtt{fluid}_5$ form a minimal feature subset.

Applying a sensitivity analysis based on Sobol indices (see Section 2.9) leads to the following five most important features (the value between brackets is a percentage explaining which variable is responsible for which fraction of the total variance of the output) for mortality prediction (SOFA): $\mathtt{age}$ (33.0%), $\mathtt{SOFA}_{\mathtt{cardio(5)}}$ (11.3%), $\mathtt{SOFA}_{\mathtt{resp(5)}}$ (10.0%), $\mathtt{SOFA}_{\mathtt{coag(5)}}$ (7.0%), $\mathtt{SOFA}_{\mathtt{cardio(3)}}$ (6.6%). This ranking corresponds approximately with the features extracted by running the backward elimination procedure. The variance of the results of the sensitivity analysis based on raw input values was too large for a correct interpretation.

Table 1
Results of mortality prediction based on the raw value dataset (above the first double line) and SOFA score dataset (below the first double line). All values represent percentages of recall r, precision p and geometric mean G, averaged over all RRSSV ($n = 100$) runs. The the improvement ΔG of using underlying raw data over using SOFA scores is shown as well. The results ($G_{A,D}$) with the same superscript symbol are not statistically different ($p \geq 0.05$, 95% confidence interval) from each other.

	SVM	ANN	RF	k-NN	ADA	Base
r_A	83.8	82.6	94.8	84.8	88.4	82.7
p_A	90.4	90.3	86.4	89.8	88.6	82.7
r_D	57.7	58.0	29.6	54.1	46.2	17.3
p_D	43.0	41.3	54.6	42.9	45.7	17.3
$G_{A,D}$	**65.9**	*65.0	60.3	*64.8	63.8	37.8
r_A	82.3	82.6	91.6	82.4	84.8	82.7
p_A	89.9	89.2	86.9	89.7	89.2	82.7
r_D	55.7	52.1	33.8	54.7	50.6	17.3
p_D	39.7	38.5	45.8	39.4	41.1	17.3
$G_{A,D}$	63.6	62.0	59.2	*63.1	*63.0	37.8
ΔG	+2.3	+3.0	+1.1	+1.7	+0.8	–

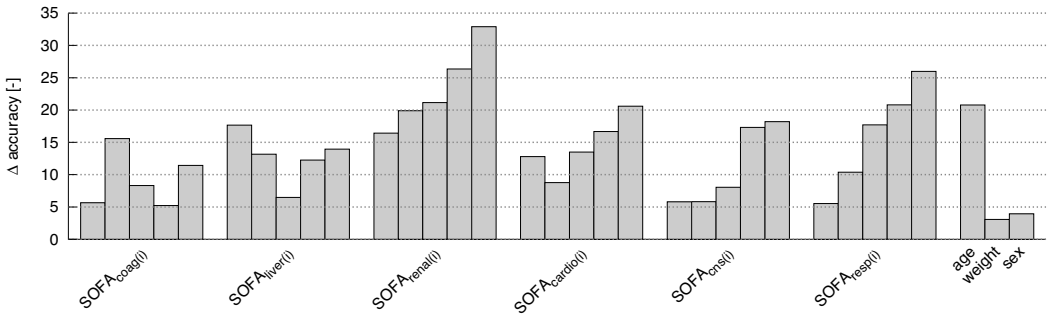
The bold values indicate the best performance metric value obtained in order to identify the best performing method.

Next, for each model, using only the most important features in combination with well-fitting parameters, the average recall r, precision p, and G value are measured, as shown in Table 1. The difference between using SOFA scores and using their underlying raw values is shown in the last row, denoted as ΔG.

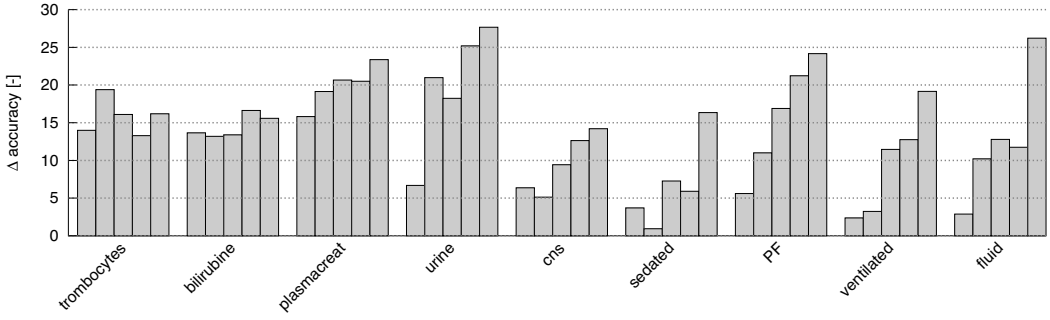
Furthermore, we apply receiver-operator curve (ROC) measurements to the different models. This can be seen in Fig. 5: Fig. 5(a) depicts the ROC measurements for the models trained on SOFA data while Fig. 5(b) shows the same measurements for models trained on raw values. The models for which no probabilistic output exists, are shown as points in the plots. For mortality prediction based on SOFA scores, the AUC for SVM is 0.769 while the AUC for ANN is 0.689. For the same prediction based on raw values, the AUC for SVM is 0.770 and the AUC for ANN is 0.689. In literature, classifiers with an AUC over 0.7 are considered acceptable [27,28].

4.3. Prolonged stay prediction

Next to modelling the patient survival, we want to make estimates regarding each individual patient LOS. First of all, a classification is made based on whether a patient will have a prolonged stay or not. To recapitulate, a prolonged stay is defined, following

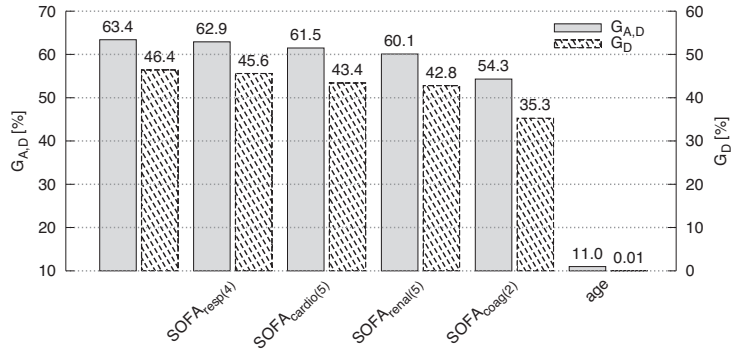


(a) Importance plot of SOFA scores

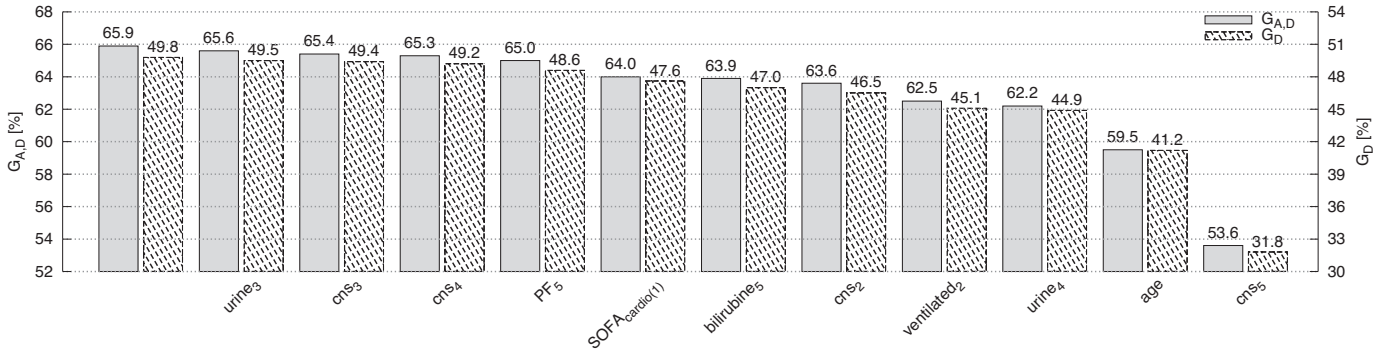


(b) Importance plot of raw values

Fig. 3. Feature importance for patient mortality classification: RF was used as a feature importance ranker (OOB error, see Section 2.9). The features are grouped per type, within each type from left to right the bars represent the feature values on day 1–5 after admission.

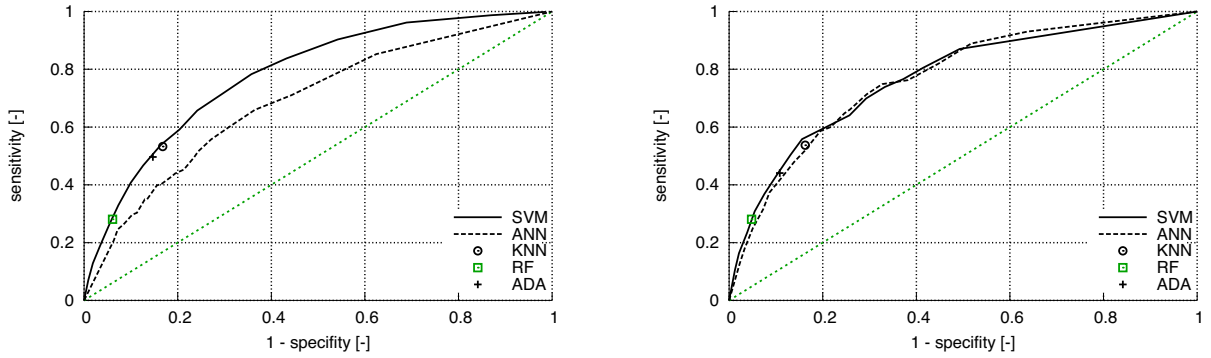


(a) Backward elimination of SOFA scores



(b) Backward elimination of raw values

Fig. 4. Backward elimination of the least informative (least decrease in $G_{A,D}$) SOFA score and raw value features. The outer-left two bars (solid and dashed, without name) always indicate the maximum $G_{A,D}$ and G_D scores obtained for the SVM model used in this test, while each subsequent bar represents the new $G_{A,D}$ and G_D scores resulting from the removal of the corresponding feature on the horizontal axis. Features are removed until only one feature remains, namely $\texttt{SOFA}_{\texttt{cns(5)}}$ in Fig. 4(a) and $\texttt{fluid}_5$ in Fig. 4(b).



(a) ROC for mortality prediction (dead) using SOFA scores. The AUC for SVM is 0.769 while the AUC for ANN is 0689.

(b) ROC for mortality prediction (dead) using raw data. The AUC for SVM is 0.786 while the AUC for ANN is 0770.

Fig. 5. ROC measurements for mortality classification models.

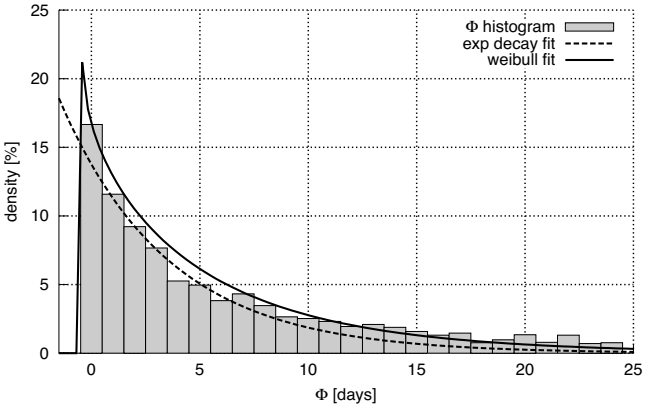


Fig. 6. Density histogram (25 bins) of Φ 5 days after admission. Furthermore, an exponential and a weibul distribution are fit on the data.

[29], as Φ > 10. A Φ histogram can be seen in Fig. 6. The probability density function (PDF) of Φ can be fitted by a weibull distribution

$$f_\Phi(\phi;k,\lambda) = \begin{cases} \frac{k}{\lambda}\left(\frac{\phi}{\lambda}\right)^{k-1} \exp -\left(\frac{\phi}{\lambda}\right)^{k} & : \phi \geq 0 \\ 0 & : \phi < 0 \end{cases}, \tag{29}$$

with $\lambda = 6.06$ (scale) and $k = 0.92$ (shape). In literature this has been shown to be a good fit for modelling LOS distributions [30], which also holds in this work.

Similar to the patient mortality modelling case, the data is distributed unevenly in terms of its classification labels. Hence re-sampling with replacement is used to redistribute the training samples of the different classes. Once again, the validation set is not re-sampled.

The correlation between $\text{SOFA}_{\text{total}}$ and Φ is shown in Fig. 7 by means of a boxplot for every $\text{SOFA}_{\text{total}}$ value. The middle box line represents the median (Q2), the lower box boundary the first quantile (Q1), the higher box boundary the third quantile (Q3), while the whiskers represent the range containing 95% of the data. This figure shows that a low $\text{SOFA}_{\text{total}}$ score indicates a high probability of having a short LOS. A higher $\text{SOFA}_{\text{total}}$ score on the other hand allows for a wider LOS range.

Each sample in the total patient dataset is labelled as short (S) (Φ < 10 days) or long (L) (Φ ≥ 10 days), indicating a prolonged stay. To obtain representative results, well-fitting parameters are sought by conducting a coarse-grained grid-search, using RRSSV ($n = 50$), in which the $G_{S,L}$ score was optimized, while restricting the model complexity, as explained in Section 4.1. Doing this resulted in the following model parameters:

- RF: 150 CART classification trees ($G_{S,L}$ converges above 150 trees).
- SVM: $C = 10$, Gaussian radial kernel with $\gamma = 0.001$; probabilistic outputs generated by fitted sigmoid function.
- ANN: 3-layer (number of layers was fixed) network with 25 hidden neurons and 1 output neuron; weight decay set to $\lambda = 0.9$.
- k-NN: classification via the mode of the $k = 150$ nearest neighbours, based on the Euclidean distance.
- ADA: 50 modified CART trees are used as weak classifiers.

Similar to the patient survival prediction experiment, the most important features are extracted by means of feature selection. The OOB importance measure based on RF is shown in Fig. 8. Here, it can once again be noticed that features based on more recent

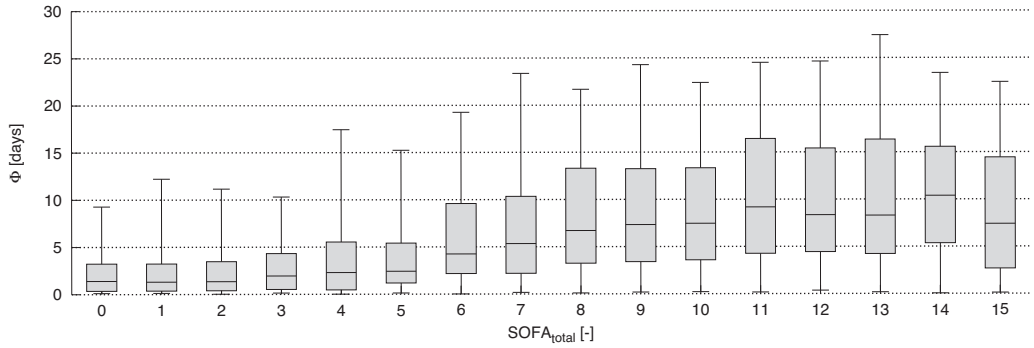


Fig. 7. Correlation between $\text{SOFA}_{\text{total}(5)}$, the total SOFA score on day 5, and Φ. $\text{SOFA}_{\text{total}(5)}$ is capped at 15, above this point data becomes too scarse to be of any relevance, which can be seen in the ECDF plot of Fig. 2.

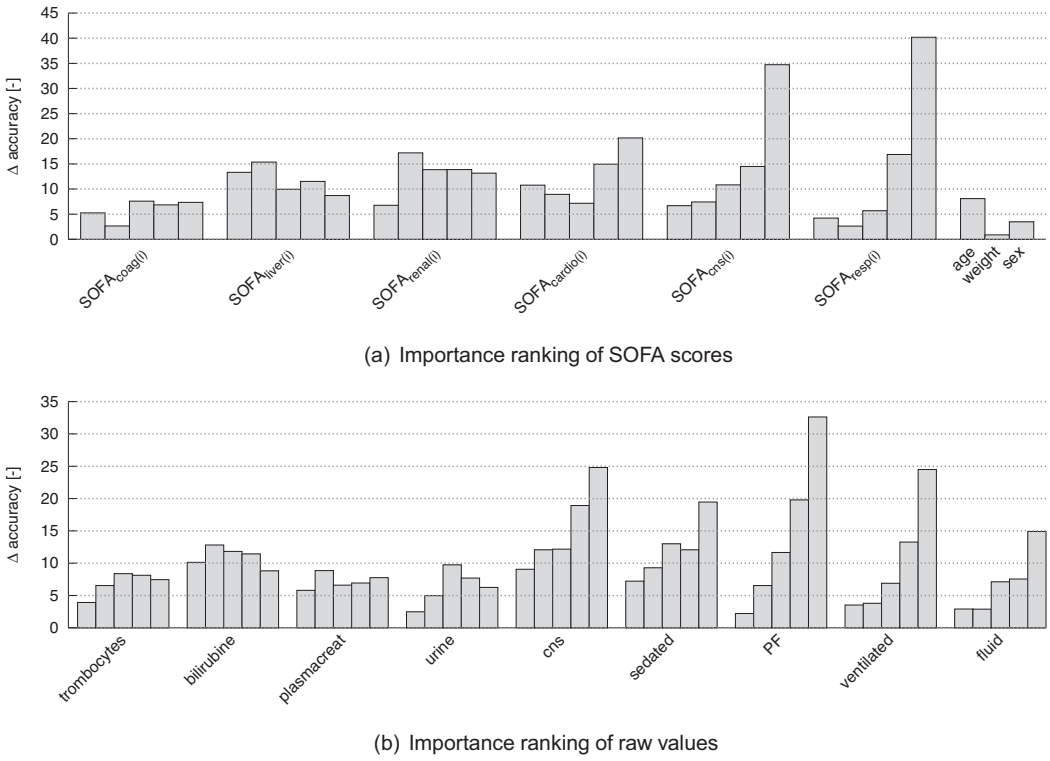


Fig. 8. Feature importance for prolonged stay classification: RF is used as a feature importance ranker (OOB error, see Section 2.9). The features are grouped per type, within each type from left to right the bars represent the value on day 1–5 after admission.

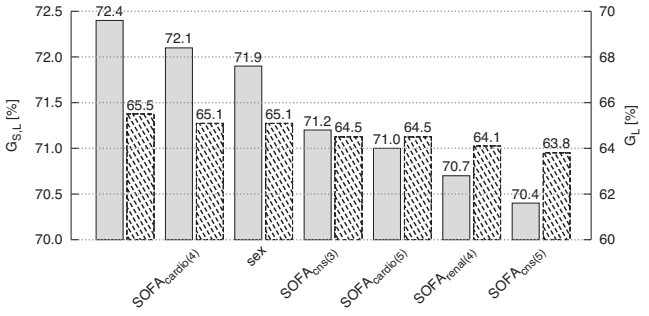


Fig. 9. Backward elimination of the least informative (least decrease in $G_{S,L}$) SOFA score features. The outer-left two bars (solid and dashed, without name) always indicate the maximum $G_{S,L}$ and G_L scores obtained for the SVM model used in this test, while each bar represents the new $G_{S,L}$ and G_L scores when removing the feature on the horizontal axis. Features are removed until only 1 feature remains, namely $\mathrm{SOFA}_{\mathrm{resp}(5)}$.

values are more important. Moreover, a large importance score for $\mathrm{SOFA}_{\mathrm{cns}(5)}$ and $\mathrm{SOFA}_{\mathrm{resp}(5)}$ are clearly noticeable. When removing raw value features sequentially by means of backward elimination, contrary to patient survival prediction, only PF_5 is deemed essential in predicting a prolonged stay. On the other hand however, using models trained only on SOFA scores, backward elimination has a slight effect on the $G_{S,L}$ value, as shown in Fig. 9. It can be noticed that $\mathrm{SOFA}_{\mathrm{cardio}(4,5)}$, sex, $\mathrm{SOFA}_{\mathrm{cns}(3)}$, $\mathrm{SOFA}_{\mathrm{renal}(4)}$, $\mathrm{SOFA}_{\mathrm{cns}(5)}$ and $\mathrm{SOFA}_{\mathrm{resp}(5)}$ form a minimal feature set for our models.

Applying a sensitivity analysis based on Sobol indices (see Section 2.9) leads to the following five most important features (the value between brackets is a percentage explaining which variable is responsible for which fraction of the total variance of the output) for prolonged stay prediction (SOFA): $\mathrm{SOFA}_{\mathrm{resp}(5)}$ (35.8%), $\mathrm{SOFA}_{\mathrm{cns}(5)}$ (27.5%), $\mathrm{SOFA}_{\mathrm{renal}(4)}$ (7.7%), $\mathrm{SOFA}_{\mathrm{cardio}(5)}$ (6.6%), $\mathrm{SOFA}_{\mathrm{cardio}(4)}$ (4.7%). This ranking corresponds approximately with the features extracted by running the backward elimination procedure. The variance of the results of the sensitivity analysis based on raw input values was too large for a correct interpretation.

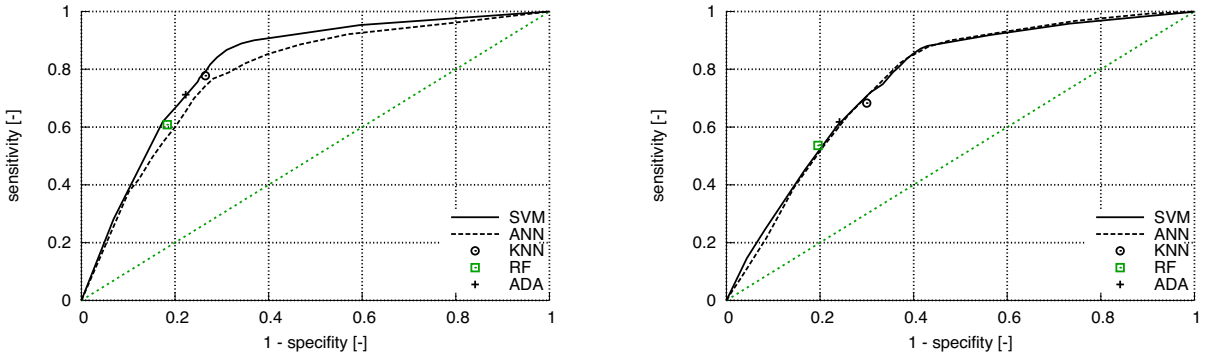
The different models, trained only on the most relevant features, using well-fitting parameters, are evaluated by measuring their recall r, precision p, and the $G_{S,L}$ measure, shown in Table 2. Contrary to the survival prediction case, in predicting a prolonged stay, the SOFA scores appear to be more predictive than the raw values.

Furthermore, we apply ROC measurements to the different models. This can be seen in Fig. 10: Fig. 10(a) depicts the ROC measurements for the models trained on SOFA scores while Fig. 10(b) shows the same measurements for models trained on raw data. The models for which no probabilistic output exists,

Table 2
Results of prolonged stay prediction using the raw data (above the first double line) and SOFA scores (below the first double line). All values represent percentages of recall r, precision p and geometric mean G, averaged over all RRSSV ($n = 100$) runs. The the improvement ΔG of using underlying raw data over using SOFA scores is shown as well. The results ($G_{S,L}$) with the same superscript symbol are not statistically different ($p \geq 0.05$, 95% confidence interval) from each other.

	SVM	ANN	RF	k-NN	ADA	Base
r_S	58.7	66.8	77.5	70.0	70.9	71.1
p_S	92.0	87.3	79.7	82.9	83.8	71.1
r_L	88.3	77.5	54.5	66.2	68.4	28.8
p_L	48.2	50.4	51.4	59.8	50.5	28.8
$G_{S,L}$	*69.2	*69.1	64.8	66.4	67.3	45.2
r_S	73.0	78.7	81.6	73.4	77.7	71.1
p_S	90.1	84.4	84.0	89.0	87.2	71.1
r_L	80.2	63.8	61.6	77.6	71.8	28.8
p_L	54.5	54.7	57.5	54.1	56.5	28.8
$G_{S,L}$	**73.2**	69.4	70.2	*72.4	*72.4	45.2
ΔG	−4.0	−0.3	−5.4	−6.0	−5.1	–

The bold values indicate the best performance metric value obtained in order to identify the best performing method.



(a) ROC for prolonged stay prediction (prolonged stay) using SOFA scores. The AUC for SVM is 0.816 while the AUC for ANN is 0.784.

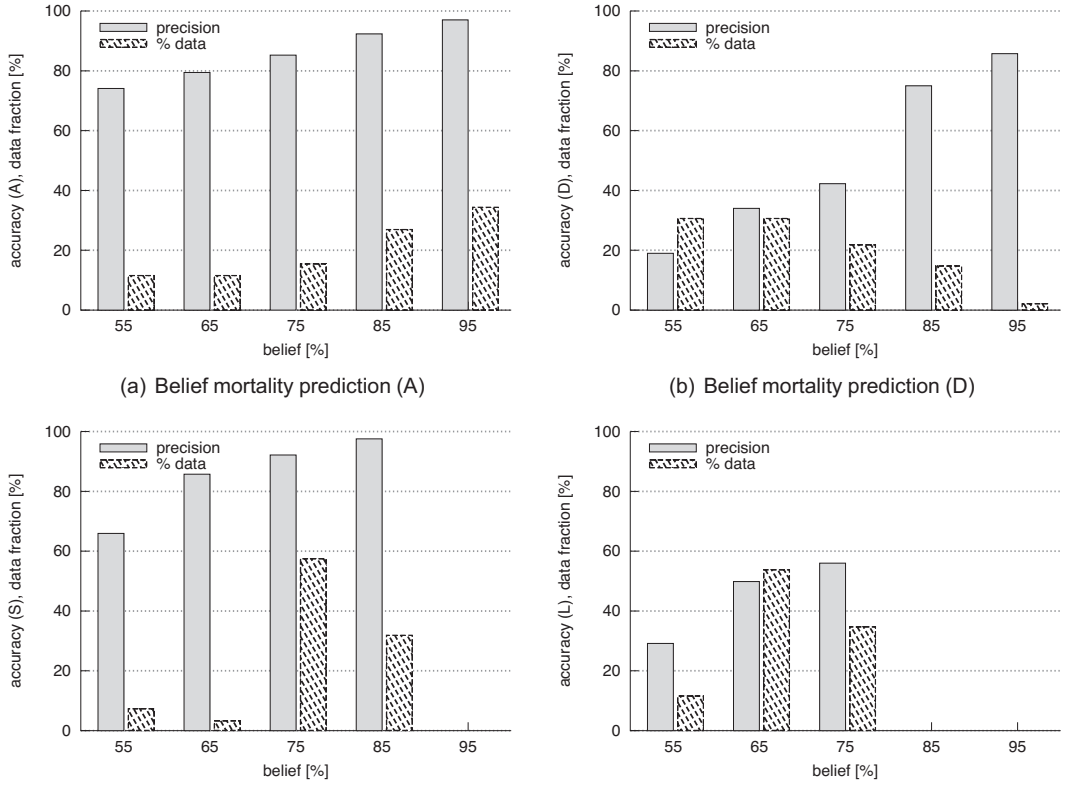
(b) ROC for prolonged stay prediction (prolonged stay) using raw data. The AUC for SVM is 0.764 while the AUC for ANN is 0.763.

Fig. 10. ROC measurements for prolonged stay classification models.

are shown as points in the plots. For prolonged stay prediction based on SOFA scores, the AUC for SVM is 0.816 while the AUC for ANN is 0.784. For the same prediction based on raw values, the AUC for SVM is 0.764 and the AUC for ANN is 0.763. In literature, classifiers with an AUC over 0.7 are considered acceptable [27,28].

4.4. Prediction belief

In order to assign a degree of belief to each classified sample, a sigmoid function is fit on the output labels of a parameter-tuned SVM. This is done both for the modelling of the patient mortality as well as prolonged stay. The sigmoid function was fit by means of a maximum likelihood (ML) algorithm, as explained in Section 2.1.1. Fig. 11 shows the probabilistic output of one validation run for both classification use cases. The classification outputs are binned per 10% belief. In this plot, the horizontal axis represents the average degree of belief the model assigns to its classification for each bin. The solid bars represent the precision obtained per belief bin, which corresponds to the true positive rate. The dashed lines represent the fraction of data that is present in each of the bins. The different sub-figures show these outputs for each possible classification label. Fig. 11(a) and (b) depict the binned precision and data fraction for the alive (A, p_A) (Fig. 11(a)) and deceased (D, p_D) (Fig. 11(b)) classification. On the other hand, Fig. 11(c) and (d) depict the precision per bin and data fraction for the nonprolonged stay (S, p_S) (Fig. 11(c)) and prolonged stay (L, p_L) (Fig. 11(d)) classification.



(a) Belief mortality prediction (A)

(b) Belief mortality prediction (D)

(c) Belief prolonged stay prediction (S)

(d) Belief prolonged stay prediction (L)

Fig. 11. Illustration of prolonged stay prediction and mortality prediction accuracy versus prediction belief (solid) using a probabilistic SVM. Furthermore, the fraction of data lying in each bin is depicted (dashed). The belief values are divided into five bins of 10% of which the middle value is shown on the horizontal axis.

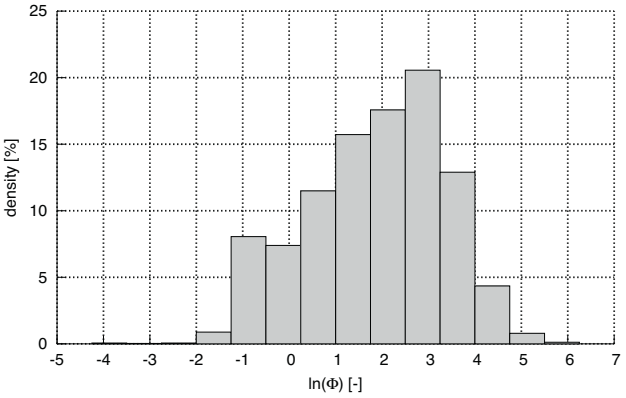


Fig. 12. Density histogram (12 bins) of logarithm of Φ.

Table 3
Two-by-two grid: data sample fraction in each cell.

	Alive (A)	Deceased (D)
short LOS (S)	*61.5% (1934)*	11.3% (355)
long LOS (L)	22.0% (693)	5.8% (181)

The italic values indicate obtaining good results for this patient set (the *61.5% (1934)*) is most relevant to ICU planning.

4.5. LOS regression

The distribution shown in Fig. 6 leads to difficulties in applying machine learning techniques due to its elongated tail. Therefore, in order to predict the numerical LOS by means of regression, it is transformed by taking its natural logarithm. This leads to a more centered distribution, as shown in Fig. 12 which depicts a density histogram (12 bins) of $\ln(\Phi)$ on the horizontal axis and the density in percentages on the vertical axis.

Outliers are recognized as patients for which LOS > 40, corresponding to the 99th percentile. These are removed from the dataset upon predicting Φ via regression analysis. Note that in the previous case of modelling patient mortality, as well as predicting whether a patient will have a prolonged stay or not, outliers were not removed.

Fig. 7 presents earlier shows the high LOS variance in patients with higher SOFA scores. Furthermore, due to the tailed distribution of the patient LOS, it is hard to accurately predict Φ for prolonged stays. Moreover, patients that do not survive their stay have a higher LOS standard deviation than patients who do survive their stay. Namely, in our patient dataset of patients with LOS ≥ 5, surviving patients have a standard deviation $\sigma = 13.4$ days while patients not surviving have $\sigma = 17.9$ days. This higher σ results from the criticality of the patient illness. To cope with this, the data has been divided into a two-by-two grid as shown in Table 3. Each cell defines a unique combination of patient mortality and prolonged stay. Besides this, this table shows the fraction of data samples belonging to each cell.

To train the different machine learning models, the training and validation datasets are also split up according to this grid. The goal is now to assess the probability of a patient belonging to the (A, D) cell. Only for this cell Φ has to be predicted accurately. This is reasonable as an accurate LOS estimate is most important for surviving patients with a nonprolonged stay. Moreover, the largest fraction of data resides in this (A, D) cell (see Table 3). Due to the data scarcity of the other cells, numerical LOS prediction might be infeasible.

The feature importance ranking of the LOS regression analysis is very similar to that of the prolonged stay prediction. Hence feature selection for LOS regression will be based on these results.

Table 4
Results of all regression methods applied to first 5 days' raw data values. The values indicate the mean absolute error $\bar{\Delta}\Phi$ and the median absolute error $\hat{\Delta}\Phi$, measured in days. The model is tested for (S, A), only to compare with models trained exclusively on SOFA scores. All results represent averages over RRSSV ($n = 100$) runs. The results with the same superscript symbol are not statistically different ($p \geq 0.05$, 95% confidence interval) from each other.

	SVR	ANN	RF	RVR	k-NN
$\bar{\Delta}\Phi$	1.77	2.12	*1.85	*1.85	*1.85
$\hat{\Delta}\Phi$	1.30	1.70	1.58	1.56	1.52

Table 5
Results of all regression methods applied to the SOFA scores of the first 5 days after admission. The values indicate the mean absolute error $\bar{\Delta}\Phi$ and the median absolute error $\hat{\Delta}\Phi$, measured in days. The data values are averages over all RRSSV ($n = 100$) runs. Additionally, the results of predicting Φ for alive patients (S, L, A) in general, and for the total data set (S, L, A, D) are shown. The results with the same superscript symbol are not statistically different ($p \geq 0.05$, 95% confidence interval) from each other.

		SVR	ANN	RF	RVR	k-NN	Base
S							
A	$\bar{\Delta}\Phi$	***1.79***	*2.03*	*1.84*	**1.86*	**1.86*	2.98
	$\hat{\Delta}\Phi$	**1.22*	*1.36*	*1.23*	***1.19***	**1.21*	2.28
D	$\bar{\Delta}\Phi$	**2.47**	3.26	2.61	2.57	2.53	3.16
	$\hat{\Delta}\Phi$	2.38	2.65	*2.26	*2.25	**2.20**	2.77
L							
A	$\bar{\Delta}\Phi$	***5.51**	5.77	5.62	***5.51**	5.55	7.73
	$\hat{\Delta}\Phi$	**4.39**	*4.70	*4.69	†4.59	†4.58	6.15
D	$\bar{\Delta}\Phi$	6.01	6.25	6.08	**5.77**	5.90	7.99
	$\hat{\Delta}\Phi$	*4.89	†5.17	†5.10	**4.65**	*4.88	6.48
(S,L)							
A	$\bar{\Delta}\Phi$	***4.41**	4.78	***4.42**	4.53	4.49	8.09
	$\hat{\Delta}\Phi$	**2.13**	2.48	2.23	*2.18	*2.19	5.44
(S,L)							
(A, D)	$\bar{\Delta}\Phi$	**4.91**	5.21	4.92	4.97	4.94	8.28
	$\hat{\Delta}\Phi$	2.64	2.87	2.60	***2.56**	***2.57**	6.32

The bold values indicate the best performance metric value obtained in order to identify the best performing method.
The italic values indicate the performance metric values for patients surviving a nonprolonged stay.

The parameters of the different models (SVR, ANN, RF, RVR and k-NN) are optimized for the average and median error, $\bar{\Delta}\Phi$ and $\hat{\Delta}\Phi$, while restricting the model complexity, using a coarse-grained grid-search with RRSSV ($n = 50$), as explained in Section 4.1. This leads to the following parameter values:

- RF: 150 CART regression trees ($\bar{\Delta}\Phi$ and $\hat{\Delta}\Phi$ converge above 150 trees).
- SVR: ϵ-SVR with $C = 10$, $\epsilon = 10^{-8}$ and a Gaussian radial kernel with $\gamma = 0.001$.
- ANN: 3-layer (number of layers was fixed) network with 10 hidden neurons and 1 output neuron; weight decay set to $\lambda = 0.9$.
- RVR: Gaussian radial kernel with $\gamma = 0.001$.
- k-NN: the average of the $k = 40$ nearest neighbours is used, based on the Euclidean distance.

In Table 5, $\bar{\Delta}\Phi$ and $\hat{\Delta}\Phi$ are shown for the different models trained on SOFA score features. The experiments are run based on RRSSV ($n = 100$) and are trained according to the grid structure defined in Table 3. The table is structured according to this previously mentioned grid: (S, A), (S, D), (L, A), and (L, D). Note that the goal is predict Φ accurately only for the two upper rows (S, A). The results for the other grid cells, (S, D), (L, A) and (L, D), are given only for comparison. Next to training our models on this grid, they are also trained solely on surviving patients (S, L, A) and on the whole

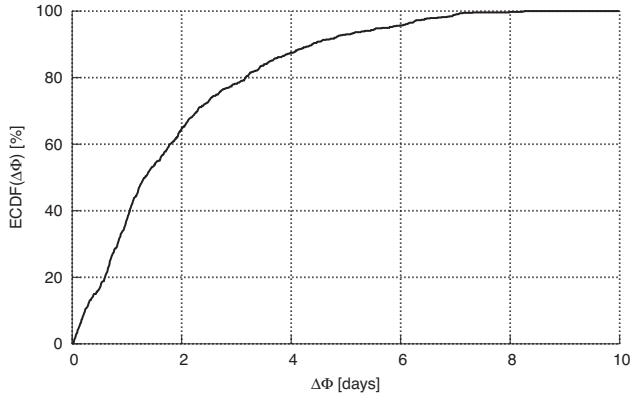


Fig. 13. ECDF plot of $\Phi_{A,S}$ for ϵ-SVR.

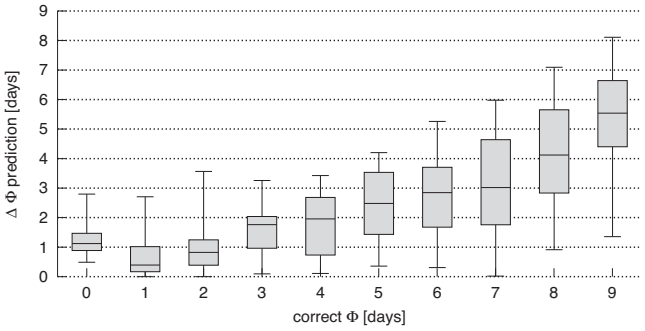


Fig. 14. Boxplot of $\Phi_{A,S}$ versus the corresponding prediction error $\Delta\Phi_{A,S}$ using an SVM. The data was grouped by rounding $\Phi_{A,S}$ to the nearest integer.

dataset (S, L, A, D). For comparison, Table 4 shows $\bar{\Delta}\Phi$ and $\hat{\Delta}\Phi$ for model training on SOFA scores for the (S, A) cell.

Fig. 13 shows an ECDF plot for the error $\Delta\Phi$. The horizontal axis shows the offset $\Delta\Phi$ in days while the vertical axis shows the ECDF value, in percentages. Fig. 14 shows a boxplot depicting the error $\Delta\Phi$ for each correct Φ-value, for the (S, A) cell. The middle line represents the median (Q2), the lower box line the first quantile (Q1), the upper box line the third quantile (Q3) and the whiskers represent the range containing 95% of the data. The data has been grouped together by rounding the exact Φ-value to the nearest integer.

4.6. Moving window patient data predictions

In the previous sections all machine learning models were trained on data values of the first 5 days after patient admission. This, however, restricts our model, preventing live predictions for each individual patient at an arbitrary point in time. As such, these models are not designed to give, for example, an accurate Φ prediction for a patient with a current LOS of 17 days. Therefore, our dataset is artificially extended via a moving window as explained in Section 3.4. To train our models, an additional feature is added, namely the current patient LOS, defined as (LOS − Φ). To train our models we have to make sure that no data from the training set spills into the validation set. Because the same patient may now be represented in two distinct samples, e.g., one sample may include patient data from day 1–5 and another sample may include data from day 13–17 although they are derived from the same patient. Though these samples have entirely different SOFA features, basic patient information, such as `age`, persists. Hence the training and validation set require disjoint patient ID sets. The models are trained with the same parameters, as well as the same minimal feature set, as obtained in the previous experiments in an RRSSV ($n = 50$) setting.

Table 6 shows the mortality prediction results of the different models by means of their precision p, recall r and $G_{A,D}$, based on moving window data. This means that for each patient an estimated outcome is given for each day he resides in the ICU. The models are trained with raw data rather than SOFA scores as this leads to better performance, which will be discussed in Section 5. Table 7 shows

Table 6
Results of moving window mortality prediction using the raw data of the previous 5 days. All values represent percentages and represent the average of recall r, precision p and geometric mean G, over all RRSSV ($n = 50$) runs. The results ($G_{A,D}$) with the same superscript symbol are not statistically different ($p \geq 0.05$, 95% confidence interval) from each other.

	SVM	ANN	RF	k-NN	ADA
r_A	82.2	80.7	92.5	81.9	81.9
p_A	87.2	86.7	82.7	86.0	87.0
r_D	59.1	57.7	34.4	54.5	58.4
p_D	49.4	46.7	57.5	46.9	48.7
$G_{A,D}$	**67.7**	65.9	62.4	65.1	67.1

The bold values indicate the best performance metric value obtained in order to identify the best performing method.

Table 7
Results of moving window prolonged stay prediction using the SOFA scores of the previous 5 days. All values represent percentages and represent the average of recall r, precision p and geometric mean G, over all RRSSV ($n = 50$) runs. The results ($G_{S,L}$) with the same superscript symbol are not statistically different ($p \geq 0.05$, 95% confidence interval) from each other.

	SVM	ANN	RF	k-NN	ADA
r_S	69.1	75.0	78.5	72.5	73.4
p_S	82.8	77.7	75.9	79.3	81.5
r_L	80.3	70.7	66.0	74.2	77.2
p_L	65.6	67.5	69.3	69.4	68.0
$G_{S,L}$	74.1	72.6	72.3	73.0	**74.8**

The bold values indicate the best performance metric value obtained in order to identify the best performing method.

Table 8
Results of all regression methods with resampling applied to moving window patient data using SOFA scores. The values indicate the mean absolute error $\bar{\Delta}\Phi$ and the median absolute error $\hat{\Delta}\Phi$, measured in days. The data values are averages over all RRSSV ($n = 50$) runs. Dashes indicate computational intractability. Additionally the results of predicting Φ for alive patients (S, L, A) in general and for the total data set (S, L, A, D) are shown. The results with the same superscript symbol are not statistically different ($p \geq 0.05$, 95% confidence interval) from each other.

		SVR	ANN	RF	RVR	k-NN
S						
A	$\bar{\Delta}\Phi$	***2.01***	*2.07*	*2.12*	–	*2.11*
	$\hat{\Delta}\Phi$	**1.57*	***1.54***	*1.56*	–	**1.58*
D	$\bar{\Delta}\Phi$	**2.35**	2.89	2.52	2.51	2.49
	$\hat{\Delta}\Phi$	2.15	2.28	**1.98**	*2.04	**2.05*
L						
A	$\bar{\Delta}\Phi$	**6.07**	6.54	6.19	–	6.32
	$\hat{\Delta}\Phi$	***4.76**	5.16	***4.78**	–	5.05
D	$\bar{\Delta}\Phi$	**6.61**	8.12	6.67	9.10	6.93
	$\hat{\Delta}\Phi$	**4.49**	6.36	5.37	7.12	5.86
(S,L)						
A	$\bar{\Delta}\Phi$	**5.44**	5.56	5.61	–	5.78
	$\hat{\Delta}\Phi$	*3.36	*3.37	**3.31**	–	3.52
(S,L)						
(A, D)	$\bar{\Delta}\Phi$	**5.99**	6.14	6.29	–	6.35
	$\hat{\Delta}\Phi$	3.91	3.84	**3.76**	–	4.00

The bold values indicate the best performance metric value obtained in order to identify the best performing method.
The italic values indicate the performance metric values for patients surviving a nonprolonged stay.

the results of predicting a prolonged stay. Contrary to the survival prediction models, these models are trained on SOFA scores, for similar reasons. The results of training modelling techniques on SOFA scores in combination with the previously mentioned grid (see Table 3) are shown in Table 8. The models are trained for each cell independently, based on training data that was also compartmentalized according to this grid. The results for the other grid cells, (S, D), (L, A), and (L, D), are given for comparison. Next to training our models on this grid, they are trained solely on alive patient data (S, L, A) and on the whole moving window dataset (S, L, A, D).

5. Discussion

In the previous section the conducted experiments and their corresponding results were set forth. Now we will discuss—following the same outline as the results section—the different interpretations of these results.

5.1. Patient mortality prediction

Regarding feature importance for our models, a general trend to be observed in Fig. 3(a) is that more recent SOFA sub-scores are more informative. This is a logical result as more recent patient information better reflects the current patient status. The same observation is made based on the raw feature importance plot in Fig. 3(b).

As explained in Section 2.9, the importance score does not take into account correlations between input features. We use backward elimination to obtain a minimal subset of features capable of attaining maximum performance. The result of the backward elimination process can be viewed in Fig. 4. In terms of SOFA score features, Fig. 4(a) shows that a small set already achieves maximum performance. In terms of raw input features this subset is slightly larger, as shown in Fig. 4(b). This relates to the SOFA score functions being an aggregation of multiple raw features. Furthermore, we see that in both sub-figures the same types of features are selected. Both select: `age`, central nervous system values ($\texttt{SOFA}_{\texttt{cns}}$ and `cns`), respiratory system values ($\texttt{SOFA}_{\texttt{resp}}$, `PF` and `ventilated`) and renal system values ($\texttt{SOFA}_{\texttt{renal}}$ and `urine`). The only difference is the selection of coagulation system values ($\texttt{SOFA}_{\texttt{coag}}$) and hepatic system values (`bilirubine`). This might be a result of different types of features holding the same information towards predicting patient mortality, which makes them interchangeable. Additionally, fluid balance scores (`fluid`) seem to be of importance in predicting via raw values, which is not available in the SOFA-based training set. The fluid balance value was not selected to be part of the SOFA score as it is not a consistent reflection of patient illness severity. Adding the fluid balance score to the SOFA dataset as an extra feature only marginally affected the results, which indicates that a lot of information represented in the fluid score is also present in the other features. Overall, these results indicate that patient mortality prediction is dependent on a plethora of organ system values. However, many features seem to be highly redundant towards the model's predictive power, which is consistent with the correlation of organ failures due to the patient illness. A general trend to be witnessed is that more recent values are better represented in the minimal feature set, which is consistent with the importance plot in Fig. 3, although less recent values are included as well. This can be explained by the fact that the model might try to find a trend in the data, which requires scores of different days.

The results of the different models are presented in Table 1. It can be immediately noticed that the p_A and r_A values (the precision and recall of predicting a patient survival) are much higher than their p_D and r_D (the precision and recall of predicting a patient death) counterparts. This likely results from, on the one hand, the data skewness towards surviving patients as the dataset only contains 8.2% deceased patients, and on the other hand, the fact that patients may be discharged from the ICU, but die afterwards in another ward. When comparing the training of models based on SOFA scores to the training based on raw values, it is noticeable that in all cases training on raw values increases the model's performance. When comparing the different models, SVMs score slightly better than the other models. They are capable of attaining $G_{A,D} = 65.9\%$. However, they are followed closely by ANN (65.0%) and k-NN (64.8%). When looking at the actual predictiveness of the SVM, we see that it is capable of getting good results for predicting surviving patients: $p_A = 90.4\%$ and $r_A = 83.8\%$. However, as explained in the previous paragraph, the results of predicting the death of a patient are worse, $p_D = 43.0\%$ and $r_D = 57.7\%$, due to the data scarcity regarding nonsurviving patients.

In summary, in case of patient mortality prediction, the SVM attains the best results. Furthermore, it could be noticed that predicting a patient survival is easier, indicated by the higher recall and precision values. In Section 5.3 it is shown how the interpretation of the SVM outputs can be enhanced by means of probabilistic outputs.

5.2. Prolonged stay prediction

Upon examining Fig. 2, we see that while having a large $\texttt{SOFA}_{\texttt{total}}$ value increases the chances of a prolonged stay, it does not rule out short stays. Hence more difficulties in predicting a prolonged stay are expected in comparison to survival prediction.

Once again, a general trend to be witnessed in the feature importance score is that more recent features hold more information regarding a prolonged stay, as shown in Fig. 8. Furthermore, there is a clear dominance of the features $\texttt{SOFA}_{\texttt{resp(5)}}$ and $\texttt{SOFA}_{\texttt{cns(5)}}$ in Fig. 8(a) and of $\texttt{PF}_5$, $\texttt{ventilated}_5$, and $\texttt{cns}_5$ in Fig. 8(b).

As explained in the previous section, this importance score does not take into account correlations between features, hence we also use backward elimination. Fig. 9 shows that, in terms of SOFA score features, $\texttt{SOFA}_{\texttt{cardio(4,5)}}$, $\texttt{SOFA}_{\texttt{cns(3,5)}}$, $\texttt{SOFA}_{\texttt{renal(4)}}$, $\texttt{SOFA}_{\texttt{resp(5)}}$, and `sex` form a minimal feature subset for our models. Noticeable is the dominant presence of more recent feature values. Running backward elimination on the raw feature values, $\texttt{PF}_5$ (which is present in $\texttt{SOFA}_{\texttt{resp}}$) forms a minimal subset singleton. This could be explained by the fact that SOFA scores form a solid aggregation of multiple raw features and actually help the model making the right decisions by its discrete if-else structure.

Inspecting the results of the different models in Table 2, generally using raw feature values decreases the model's performance. Although we would expect information loss by the SOFA score aggregation function, this could be related to increased feature expressiveness by aggregating them into SOFA scores, which is consistent with the explanation in the previous paragraph showing that $\texttt{PF}_5$ forms a minimal subset singleton. The best performing model appears to be once again an SVM, with $G_{S,L} = 73.2\%$, followed by k-NN (72.4%) and ADA (72.4%). The difference in recall between prolonged stays ($r_L = 80.2\%$) and nonprolonged stays ($r_S = 73.0\%$) appears to be much smaller than the difference in precision ($p_S = 90.1\%$ and $p_L = 54.5\%$). Thus, the precision of predicting nonprolonged stays seems to be higher than predicting prolonged stays. This is consistent with Fig. 7 in which a low SOFA score indicates a short stay with low variance. When looking at high SOFA scores, we see that the LOS variance is much higher, resulting in a lower prediction precision.

In summary, similar to patient mortality prediction the SVM outputs the best results. Furthermore, the prediction precision of nonprolonged stays is higher than the precision of prolonged stays. Similar to survival modelling, the interpretation of the SVM

outputs can be enhanced by assigning probabilities to the outputs, as will be shown in the following subsection.

5.3. Prediction belief

It is favorable that the models used for classification assign a degree of belief to their output. For this reason, the best performing model, namely the SVM, has been extended by fitting a sigmoid function via maximum likelihood to its binary output.

In Fig. 11 an illustration of this belief assignment is shown together with the classification precision, as well as the data fraction belonging to each belief bin. This figure indicates that the belief assigned by the model positively correlates with the model precision. A higher belief accords to a higher precision. In Fig. 11(a) it can be noticed that, even at low belief values, the model precision regarding the prediction of patient survival (A) is high. Moreover, the data appears to be centered around higher belief values indicating that the model is very certain in predicting patient survivals (largest fraction of data is in the 95% belief bin). Fig. 11(b) shows the precision of the prediction of a patient death (D). Low belief values have a minimal precision, making their predictions highly uncertain. The data distribution is also left-centered, indicating the model's inability to accurately predict patient deaths. Similar to Section 5.1, where the prediction performance of predicting a patient's death is much lower than predicting a patient's survival, the overall certainty in death predictions is much lower. Only for a minority of patients the model is very certain about their death, whereas in terms of patient survival prediction the model is certain for the majority of patients.

Fig. 11(c) shows the prediction belief of a nonprolonged patient stay (S). The main data fraction seems to be located around a belief of 75%, while none of the data belongs to the 95% belief bin. However, it should be noted that the precision of the output samples belonging to the 85% belief bin is approximately the same as the 95% belief bin in Fig. 11(a). Looking at the prediction of prolonged stays (L) in Fig. 11(d), no outputs are assigned a 85% or 95% belief. Moreover, the overall precision appears to be quite low. This is, similar to the explanation of the previous paragraph, consistent with the results of Section 5.2 in which the precision of predicting a prolonged stay is much lower than the precision of predicting a nonprolonged stay.

In summary, the model's belief does not have a one-to-one mapping with its precision, however, higher beliefs corresponds to a higher precisions. Physicians can thus use this belief to alter their interpretation of the SVM outputs. As such, it is possible to only take into account outputs surpassing a pre-defined belief threshold, which could be defined via cross-validation. Using the illustration in Fig. 11(b), we could for example assign a threshold of 80% belief. As such, physicians would discard any results lower than this threshold as these are inherently uncertain. However, in case of a prediction belief of over 80% (corresponding to an accuracy of over 70%), more faith could be put in the classification output.

5.4. LOS regression

Table 3 already showed that the patient data is highly skewed towards short stays and patients who survive, indicated by the high fraction of data (61.5%) belonging to the (S, A) cell. Our goal is to predict the remaining LOS Φ exactly only for those patients belonging to this particular cell. This is done for two reasons, (i) modelling the LOS for the other cells is difficult due to their large LOS variance, and (ii) predicting the mortality and LOS for this group of patients is of most importance to ICU resource planning, which is generally done only a few days in advance.

The results of the different regression models are shown in Table 5. As explained in the previous paragraph, the first two rows, corresponding to the (S, A) cell, are of the most interest. All models seem to be equally predictive, except for ANNs, with approximately $\bar{\Delta}\Phi = 1.79$ days and $\hat{\Delta}\Phi = 1.22$ days. Fig. 13 shows an ECDF plot for this cell. This plot highlights the low error rate for a large fraction of the total patient dataset, indicated by the steep slope at the beginning of the curve. This is related to the low $\hat{\Delta}\Phi$ value in Table 5. As such, about 70% of the total patient dataset that remained longer than 5 days in the ICU have an error rate of less than two days. Fig. 14 shows these results from a different angle. Here the correct LOS values are rounded to the nearest integer and boxplots are drawn according to the corresponding error rates. This clearly shows that the error rate scales with the correct LOS, leading to low median errors for predictions of short stays. This is good news as we are able to predict short stays very accurately, while most of the error concentrates around longer stays. These results are entirely based on models trained on SOFA score features, which are selected according to the backward eliminiation minimal subset presented in Section 5.2. For comparison's sake, we also trained the models on raw feature values for the (S, A) cell, for which the results are shown in Table 4. The average $\bar{\Delta}\Phi$ is approximately the same for training based on SOFA score features as for training based on raw data features, however, in general the median $\hat{\Delta}\Phi$ increases. This reinforces our belief that using SOFA score features is indeed an improvement over using their raw underlying values regarding the LOS predictions of our models.

Additionally, rows 2 and 3 in Table 5 show the results when training the model on data corresponding to the (S, D) cell. The higher average and median values (for SVR, $\bar{\Delta}\Phi = 2.47$ days versus $\bar{\Delta}\Phi = 1.79$ days and $\hat{\Delta}\Phi = 2.38$ days versus $\hat{\Delta}\Phi = 1.22$ days) indicate the increased variance of the LOS of patients that do not survive their ICU stay. Looking at the results in case the models are trained on prolonged LOS data, we see that the $\bar{\Delta}\Phi$ and $\hat{\Delta}\Phi$ values become very large (for SVR, $\bar{\Delta}\Phi = 5.51$ days and $\hat{\Delta}\Phi = 4.39$ for patients who survive and $\bar{\Delta}\Phi = 6.01$ days and $\hat{\Delta}\Phi = 4.89$ for patients who die). An explanation is that on the one hand it is hard to predict a long LOS in advance, due to the larger LOS variance for higher SOFA scores as explained in Section 5.2, and on the other hand that data is very scarce for prolonged stays. Furthermore, the models are trained exclusively on data of patients surviving their stay, corresponding to the (S, L, A) row. Here $\bar{\Delta}\Phi$ and $\hat{\Delta}\Phi$ are both very high (for SVR, $\bar{\Delta}\Phi = 4.41$ days and $\hat{\Delta}\Phi = 2.13$ days) compared to the (S, A) case, indicating that the prolonged stays severely deteriorate the model's performance. Even worse results are obtained when training on the whole dataset, as shown in the bottom two rows (for SVR, $\bar{\Delta}\Phi = 4.91$ days and $\hat{\Delta}\Phi = 2.64$ days). These results indicate that splitting the data in a two-by-two grid improves the model's accuracy where it is most relevant.

Furthermore, we compare our results with those of Kramer and Zimmerman [10], who also predicted Φ based on measured patient values of the first 5 days. Because they removed patients with LOS > 30 from the dataset, we do the same to obtain a good comparison. Their models are evaluated by measuring the r^2 value of the regression function. While they report an r^2 value of 18.2% over all individual patients, our best-performing method, SVR, attains an r^2 value of 21.9% in an RRSSV ($n = 100$) setting. This indicates that our technique outperforms their model regarding predicting individual patient LOS.

In summary, a two-by-two grid was developed which acts as a classification pre-processor. This grid classifies patients into a cell according to their mortality and prolonged LOS predictions. Only for the cell in which patients survive their stay and have a nonprolonged stay, LOS regression is accurate. This also coincides with the most important patient group in terms of ICU resource planning. SVR trained on SOFA scores delivered the best results, in which patients of this previously mentioned group can be predicted with an average error of 1.79 days and a median error of 1.22 days. When

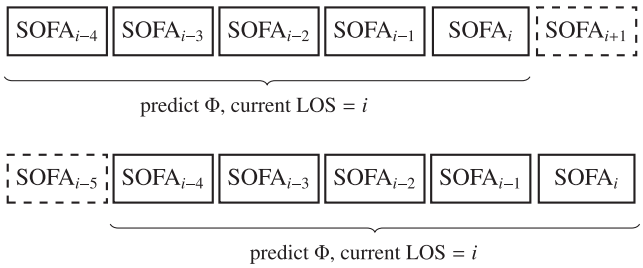


Fig. 15. Illustration: moving data window for SOFA scores. SOFA score feature values from the previous 5 days are used to predict Φ; for the moving window $i \in \{5, \ldots, \text{LOS}\}$, while in the original nonmoving case $i = 5$.

looking at the error-vs-LOS plot, we see that the error scales with the LOS, indicating a highly accurate LOS prediction for patients with short stays.

5.5. Moving window patient data prediction

Our models are also evaluated by training them on a moving data window of 5 days, as illustrated in Fig. 15. This allows us to predict the mortality, prolonged LOS and remaining LOS for patients with an arbitrary current LOS, instead of only patients with a LOS = 5. Based on Tables 6 and 7, we see approximately the same results when using moving window data as when using nonmoving window data (thus only for patients having a current LOS of 5 days; predictions based on the first 5 days). These results are obtained by using the same set of features as used for training the models based on nonmoving window data.

However, looking at the LOS regression results for moving window data in Table 8, we see that the results are overall slightly worse than predicting based on nonmoving window data with approximately $\bar{\Delta}\Phi = 2.01$ days and $\hat{\Delta}\Phi = 1.57$ days, for the (S, A) cell. Therefore, it is slightly harder to predict Φ for a patient on an arbitrary point (later than 5 days after admission) in time. This could be due to the model's confusion as it now receives data inputs of different timestamps. For example, following the illustration in Fig. 15, when $i = 7$, $\text{SOFA}_{i-4} = \text{SOFA}_3$ acts as the same input feature as SOFA_1, when predicting a patient with a current LOS of 7 days instead of a current LOS of 5 days.

In summary, the predictions of a prolonged stay and mortality are approximately the same for moving and nonmoving data. However, the LOS regression results are slightly worse for moving data, making it slightly harder to predict Φ for patients at an arbitrary point in time. Nonetheless, the results are sufficiently good ($\bar{\Delta}\Phi = 2.01$ days and $\hat{\Delta}\Phi = 1.57$ days) for applying our models in a moving data setting, which allows for live patient predictions.

6. Conclusion

In this work, different machine learning models were applied, tuned and evaluated to predict the individual patient mortality, length of stay (LOS) and prolonged stay in the intensive care unit (ICU), using a 14,480 patient dataset. Each model was trained both on raw data values as well as sequential organ failure (SOFA) scores of the first five days after admission. Our models are capable of attaining geometric mean accuracy values of $G_{A,D} = 65.9\%$ (AUC of 0.77) and $G_{S,L} = 73.2\%$ (AUC of 0.82) for predicting patient mortality and predicting a prolonged stay. Interpretation of the model outputs was enhanced by fitting a probabilistic model. As such, the prediction accuracy has been characterized as a function of a belief percentage. This allows ICU intensivists to assess the credibility of the predictions, and act accordingly. Training models on the whole dataset for LOS regression appeared to be infeasible, due to the high LOS variance of deceased patients and patients with a prolonged stay. Therefore, a two-by-two grid was constructed based on the output of the prolonged stay and mortality classifiers. By training models on only a subset of the total dataset, we are able to attain average and median error rates of 1.79 and 1.22 days, in predicting the remaining LOS for surviving patients with a nonprolonged stay. We have shown that our models are able to perform approximately equally well when using a moving data window. This indicates their applicability in a real-time ICU monitoring environment.

Future work will encompass the comparison of our model predictions with predictions made by ICU physicians. Furthermore, we will model the ICU load by means of survival analysis. In a later stadium we would like to implement our models in a real ICU environment, generating live patient analyses in order to assist physicians in assessing the current and future patient status.

References

[1] Rapoport J, Teres D, Zhao Y, Lemeshow S. Length of stay data as a guide to hospital economic performance for ICU patients. Med Care 2003;41(3):386–97, http://dx.doi.org/10.1097/01.mlr.0000053021.93198.96.

[2] Vincent JL, de Mendonca A, Cantraine F, Moreno R, Takala J, Suter PM, et al. Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study. Crit Care Med 1998;26(11):1793–800, http://dx.doi.org/10.1097/00003246-199811000-00016.

[3] Moreno R, Vincent JL, Matos R, Mendonca A, Cantraine F, Thijs L, et al. The use of maximum SOFA score to quantify organ dysfunction/failure in intensive care. results of a prospective, multicentre study. Intensive Care Med 1999;25(7):686–96, http://dx.doi.org/10.1007/s001340050931.

[4] de Mendona A, Vincent JL, Suter PM, Moreno R, Dearden NM, Antonelli M, et al. Acute renal failure in the ICU: risk factors and outcome evaluated by the SOFA score. Intensive Care Med 2000;26(7):915–21, http://dx.doi.org/10.1007/s001340051281.

[5] Ferreira F, Bota D, Bross A, Mélot C, Vincent J. Serial evaluation of the SOFA score to predict outcome in critically ill patients. J Am Med Assoc 2001;286(14):1754–8, http://dx.doi.org/10.1001/jama.286.14.1754.

[6] Regel G, Grotz M, Weltner T, Sturm JA, Tscherne H. Pattern of organ failure following severe trauma. World J Surg 1996;20(4):422–9, http://dx.doi.org/10.1007/s002689900067.

[7] Antonelli M, Moreno R, Vincent J, Sprung C, Mendoça A, Passariello M, et al. Application of SOFA score to trauma patients. Sequential organ failure assessment. Intensive Care Med 1999;25(4):389–94.

[8] Zimmerman JE, Kramer AA, McNair DS, Malila FM, Shaffer VL. Intensive care unit length of stay: benchmarking based on acute physiology and chronic health evaluation (APACHE) IV. Crit Care Med 2006;34(10):2517–29, http://dx.doi.org/10.1097/01.CCM.0000240233.01711.D9.

[9] Verduijn M, Peek N, Voorbraak F, de Jonge E, de Mol B. Dichotomization of ICU length of stay based on model calibration. In: Proceedings of the 10th Conference on Artificial Intelligence in Medicine. AIME'05;. 2005. p. 67–76, http://dx.doi.org/10.1007/11527770_10.

[10] Kramer AA, Zimmerman JE. A predictive model for the early identification of patients at risk for a prolonged intensive care unit length of stay. BMC Med Inform Decis Mak 2010;10(27), http://dx.doi.org/10.1186/1472-6947-10-27.

[11] Sandri M, Berchialla P, Baldi I, Gregori D, De Biasi R. Dynamic Bayesian networks to predict sequences of organ failures in patients admitted to ICU. J Biomed Inform 2013;48:106–13, http://dx.doi.org/10.1016/j.jbi.2013.12.008.

[12] Meyfroidt G, Guiza F, Cottem D, De Becker W, Van Loon K, Aerts JM, et al. Computerized prediction of intensive care unit discharge after cardiac surgery: development and validation of a Gaussian processes model. BMC Med Inform Decis Mak 2011;11(64), http://dx.doi.org/10.1186/1472-6947-11-64.

[13] Silva A, Cortez P, Santos MF, Gomes L, Neves J. Rating organ failure via adverse events using data mining in the intensive care unit. Artif Intell Med 2008;43(3):179–93, http://dx.doi.org/10.1016/j.artmed.2008.03.010.

[14] Cell LA, Galvin S, Davidzon G, Lee J, Scott D, Mark R. A database-driven decision support system: customized mortality prediction. J Pers Med 2012;2(4):138–48, http://dx.doi.org/10.3390/jpm2040138.

[15] Bishop CM. Neural networks for pattern recognition. 1st ed. USA: Oxford University Press; 1996. ISBN: 0198538642.

[16] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat 1992;46(3):175–85, http://dx.doi.org/10.2307/2685209.

[17] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273–97, http://dx.doi.org/10.1023/A:1022627411411.

[18] Breiman L, Friedman J, Olshen R, Stone C. Classification and regression trees. 1st ed. Wadsworth and Brooks; 1984, http://dx.doi.org/10.2307/2530946. ISBN: 0412048418.

[19] Breiman L. Random forests. Mach Learn 2001;45(1):5–32, http://dx.doi.org/10.1023/A:1010933404324.

[20] Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting. Ann Stat 2000;28(2):337–407, http://dx.doi.org/10.1214/aos/1016218223.
[21] Tipping ME. Sparse Bayesian learning and the relevance vector machine. J Mach Learn Res 2001;1:211–44.
[22] Gorissen D, Crombecq K, Couckuyt I, Demeester P, Dhaene T. A surrogate modeling and adaptive sampling toolbox for computer based design. J Mach Learn Res 2010;11:2051–5.
[23] Platt JC. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in large-margin classifiers. MIT Press; 1999. p. 61–74.
[24] Sobol IM. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. Math Comput Simul 2001;55(1-3):271–80, http://dx.doi.org/10.1016/S0378-4754(00)00270-6.
[25] Cortez P, Embrechts MJ. Using sensitivity analysis and visualization techniques to open black box data mining models. Inform Sci 2013;225:1–17, http://dx.doi.org/10.1016/j.ins.2012.10.039.
[26] Li DC, Fang YH, Fang YMF. The data complexity index to construct an efficient cross-validation method. Decis Support Syst 2010;50(1):93–102, http://dx.doi.org/10.1016/j.dss.2010.07.005.
[27] Rosenberg AL. Recent innovations in intensive care unit risk-prediction models. Curr Opin Crit Care 2002;8(4):321–30, http://dx.doi.org/10.1097/00075198-200208000-00009.
[28] Ohno-Machado L, Resnic FS, Matheny ME. Prognosis in critical care. Annu Rev Biomed Eng 2006;8:567–99, http://dx.doi.org/10.1146/annurev.bioeng.8.061505.095842.
[29] Laupland KB, Kirkpatrick AW, Kortbeek JB, Zuege DJ. Long-term mortality outcome associated with prolonged admission to the ICU. CHEST J 2006;129(4):954–9, http://dx.doi.org/10.1378/chest.129.4.954.
[30] Marazzi A, Paccaud F, Ruffieux C, Beguin C. Fitting the distributions of length of stay by parametric models. Med Care 1998;36(6):915–27, http://dx.doi.org/10.1097/00005650-199806000-00014.