On Risk Stratification Strategies in Intensive Care Medicine

Martin MacGuill* Tobias Petri* Ralf Zimmer

August 19, 2011

LMU Munich, Department of Informatics
Amalienstraße 17
80333 Munich, Germany
+49 89 2180 4042
http://www.bio.ifi.lmu.de
{macguill,petri,zimmer}@bio.ifi.lmu.de

[1] authors contributed equally

Abstract

Outcome in critically ill patients is thought to be dependent upon multiple parameters that interact in a complex manner. Experienced critical care doctors are capable of making reasonably accurate prognoses however variation between prognoses is very high. Objective risk estimation procedures help to overcome these differences. With advances in automated patient surveillance there is increasing potential for more rapid and accurate prognostication by means of computational models. We show that state-of-the-art machine learning techniques employed on real world intensive care unit data are on par with conventional risk estimation models developed on pristine research data sets. We set up a rigorous validation environment to estimate objective performance values of both conventional and novel risk and mortality models. Our results show that multiple performance measures must be taken into account when assessing the value of a model. With regard to application into routine medical practice, no single classifier is superior. Careful definition of the model characteristics most desirable to health care providers are essential before choosing one risk model over another.

1 Introduction

Highly accurate patient-specific prediction of outcome would revolutionise practice in critical care medicine. Since the early 1980s numerous computational methods have achieved varying degrees of success regarding incorporation into routine practice in intensive care units (ICUs). The most widely used methods are regression-based risk stratifiers (RS) that are based on physiological variables. Examples include the Acute Physiology and Chronic Health Evaluation (APACHE) model [19], the Mortality Prediction Model (MPM, [23]), and the Simplified Acute Physiology Score (SAPS, [22]). The development and testing of these models has firmly established a connection between prognosis and physiology-based variables in ICU patients.

Risk stratifiers have been shown to be of real value in the following areas: improving the quality of care being provided to patients [33], controlling for variations in severity of illness between ICUs and hospitals when performing audit or allocating resources [14], and when selecting subjects for participation in clinical trials.

There is broad consensus supporting the use of risk stratification for these purposes however using them as a basis for clinical decision making on an individual patient basis is not appropriate. In this work we investigate state-of-the-art machine learning methods as risk stratifiers. In particular we apply support vector machines, decision trees and random forest approaches. A novel database (MIMICII, [36]) is used as the data source enabling our models to integrate time series markers. We incorporate information generated during the first 48 hours of admission and so our models can be said to be learning from the clinical development of a patient over time.

2 Related Work

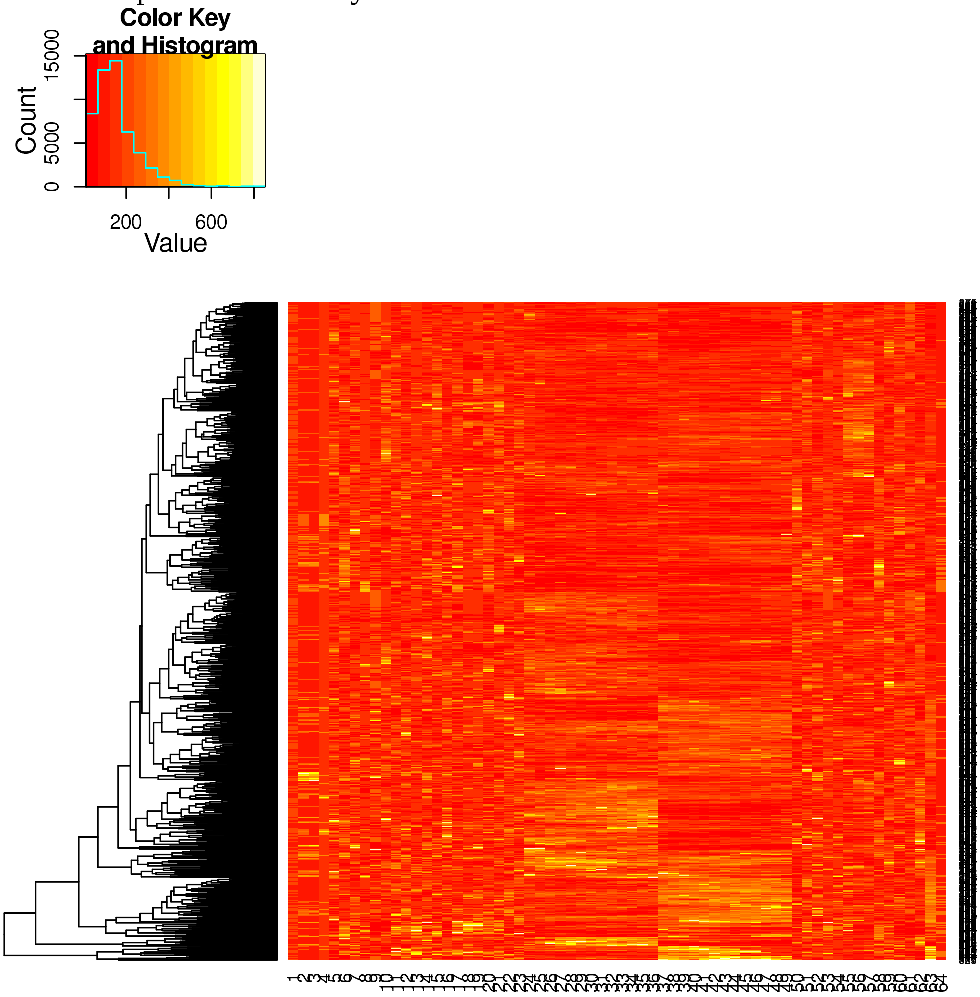
To date all the widely used risk stratification models (RS) in intensive care medicine are based on logistic regression. To faciliate usage by health care providers these quite complex models are generally further reduced to simpler scoring schemes.

Examples of risk stratifiers are the Mortality Prediction Model (MPM, [23]) and the Simplified Acute Physiology Score (SAPS, SAPSII – [22]). The Acute Physiology and Chronic Health Evaluation scores are the most widely used and their history serves to illustrate some

Table 1: **AUROC values of mortality models.** This table shows an overview of several studies that compare the most widely used mortality prediction models in intensive care medicine. The studies show overall good discrimination (as represented by areas under the ROC curves) but often poor calibration, specifically overprediction of mortality [38]. MPM0:=Mortality Probability Model, MPM/-II24:= Mortality Prediction Model-24 hours, SAPS:=Simplified Acute Physiology Score, AP:=Acute Physiology and Chronic Health Evaluation II, (a=full set, b=validation set).

Author	**AP-II**	**MPM0**	**MPM24**	**SAPS**	**AP-III**	**SAPS-II**	**MPM-II0**	**MPM-II24**
Castella [5]	0.867	0.865	-	-	-	-	-	-
Rowan [34]	0.83	0.74	-	-	-	-	-	-
Wilairatana [43]	0.723	-	-	0.71	0.694	-	-	-
Del Bufalo [2]	0.808	-	-	-	-	0.735	-	-
Castella (a) [4]	0.852	0.773	0.825	0.798	0.866	-	-	-
Castella (b) [4]	0.857	0.778	0.815	0.799	-	0.855	0.815	0.833
Moreno [29]	-	-	-	-	-	0.822	0.785	-
Nouira [31]	0.82	-	-	-	-	0.84	0.85	0.882
Tan [38]	0.88	-	-	-	-	0.87	-	-
Patel [32]	0.702	-	-	-	-	0.672	-	0.695
Vassar [40]	0.87	-	-	-	0.89	-	-	-
Katsaragakis [18]	0.839	-	-	-	-	0.87	-	-
Livingston [26]	0.763	-	-	-	0.795	0.784	0.741	0.791
Capuzzo [3]	0.805	-	-	-	-	0.816	-	-
Markgraf [28]	0.832	-	-	-	0.846	0.846	-	-
Beck [1]	0.835	-	-	-	0.867	0.852	-	-

Figure 1: Heatmap of standardized values showing a patient clustering of all 64 parameters. The last column represents mortality.

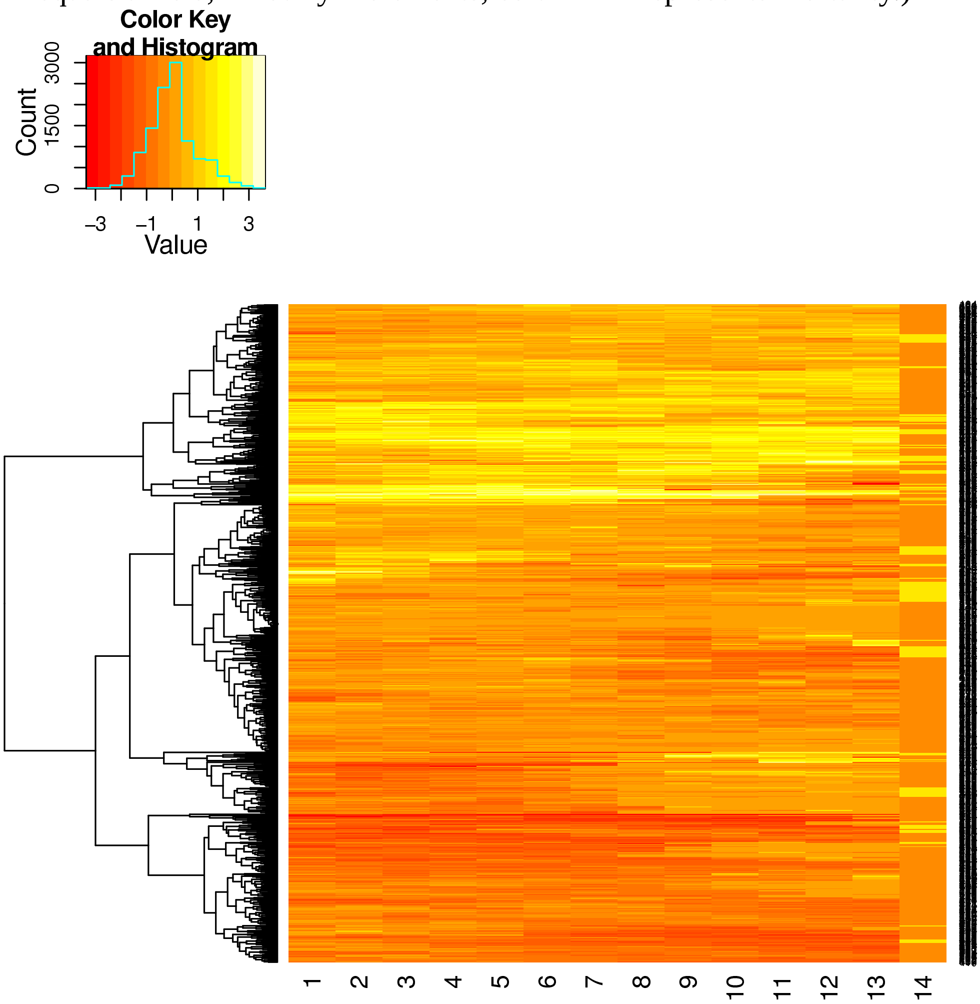


of the problems in this field. The original APACHE score was the first risk model based on physiological variables and was developed in 1981 [21]. The physiological parameters (the independent variables in the logistic regression) were selected by an expert panel of doctors. The follow-up APACHE II model was published in 1985 [20] and had a reported AUC of 0.86 on an evaluation set. APACHE III never became widely used because the authors decided to make the system proprietary.

Although there are examples of risk stratifiers being controversially used to aid clinical decisions such as whether admission to the ICU is futile or whether to end therapy [8, 16] there is very broad consensus that currently available risk stratifiers are at best useful for controlling for variations in severity of illness of patients between ICUs and between hospitals [33]. It is accepted that their capacity to predict outcome on an individual patient basis is very limited. This is due to problems of both calibration and discrimination. The predictive logistic models are usually calibrated to fit observed risks for, for instance, a specific ward or a specific hospital and accordingly they tend not to generalize well. A well-calibrated model is a model that generalizes when applied to novel data without loosing its predictive power.

A number of studies have compared performance between the best known risk stratification models. The area under the ROC (AUROC) has generally been used to compare models and the published AUCs for a number of different models are shown in Table 1. Using the AUROC as a basis for model selection has been criticized [7] due to the fact that its use can lead to over-

Figure 2: Heatmap of standardized values showing a patient clustering of heart rate timelines (column 1 equals time 0, 4-hourly increments, column 14 represents mortality.)



fitting and over-sensitivity. As can be seen in the table, while each new model had improved discrimination when compared to its predecessors, no model is clearly superior.

We are now in an age of automated generation of large volumes of intensive care unit data, so it seems a natural progression to employ machine learning methods. A range of machine learning techniques have been used in the critical care setting and described in the medical literature [10, 12, 15, 24, 25, 27, 30, 41]. Machine learing techniques demonstrate comparable discrimination with logistic regression but have not as yet been conclusively shown to be superior.

As a source of data, the MIMICII database [36] offers well-structured time-stamped patient data. It has for example been used to find risk factors for the acute respiratory distress syndrome [17] or to show that certain ICU practices varied significantly as a function of time of day (i.e. care provided at night is different from that provided during the day [35]).

In this work, we place the problem of risk stratification into a machine learning setting. We focus on the development of a patient specific predictive risk model. Reported cross validation performance as well as scores on an independent validation set provide class specific sensitivity and specificity values. By using SAPSI scores as baseline values, a direct comparison and rigorous assessment of several methods was feasible.

3 Data

The MIMIC II Database. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) is a publicly available database of intensive care unit patient data. It is a substantial and very comprehensive database containing anonymized demographic, clinical (such as admission diagnoses) and physiological data, laboratory results, detailed documentation of treatment and free text records [36]. One of the major strengths of the database is that it offers high temporal resolution for certain parameters such as heart rate, blood pressure, oxygen saturation and respiratory rate. In summary, the database is composed of 25,328 ICU patient records. The median (interquartile range) ICU stay is 2.2 (1.1-4.4) days and the overall mortality rate is 11.7%. We extracted data for the following parameters: number of hospital admissions, number of ICU admissions, gender, SAPSI score, SOFA score, creatinine, partial pressure of CO2 in blood, bilirubin, arterial blood pH, white cell count, respiratory rate, lactic acid, glucose, potassium, sodium, coagulation, ventilation parameters. temperature, heart rate, blood pressure, body weight, diagnoses, catecholamine doses, volume of red cell concentrate infused, total fluid input and output, urea, hematocrit, bicarbonate and Glasgow coma scale.

4 Methods

4.1 Model and Data Representation

The aim of this work is the prediction of a binary outcome: mortality. Using machine learning terminology, we refer to these outcomes as labels $l \in \{0, 1\}$.

An abstraction of the patient data is given by a matrix $F \in \mathcal{F}^{n \times m}$ of feature values representing each of n patients in a row i by m features $F_i \in \mathcal{F}^m$. A feature can be numeric, binary or nominal. Each row F_i in the matrix is associated with an outcome l_i, the label. For the prediction of l_i we train a model $M : \mathcal{F}^m \rightarrow L, M(F_i) = l_i$. Since we are dealing with missing values – round the clock monitoring of an intensive care patient without human or measurement failure is not realistic – we use a column median replacement strategy. Each missing $f_{ij} \in F$ is replaced by $median(f_{1j}, ..., f_{nj})$. If more than 30% of all features are missing the patient is not allocated to any set.

4.2 Cross Validation

Predictive methods like regression, decision trees or random forests must be trained and tested on two independent sets to avoid overfitting. Most preceeding studies apply a single 2 to 1 split into a training set (the derivation group) and a test set (validation group). To rigorously compare several methods we use a k-fold stratified cross validation (k-CV) and recompute risk models for each fold: the rows of the feature matrix F are shuffled and the patient set is divided into k equally sized sets with equal label distribution. Given the joint CV or validation set predictions we can now compute test statistics for the assessment of multiple methods. Additionally, we compile a completely separate validation set to assess the model quality of a method trained on the complete training set.

4.3 Test Statistics and Scoring

Given a prediction vector $l' \in L^n$ of all patients and the known labels $l \in L^n$ we compute several different statistics. For the binary outcomes, we count true negatives (TN), false negatives (FN), true positives (TP) and false positives (FP). Derived measures are the true positive rate (TPR) defined as TP/(TP+FN) and the true negative rate (TNR) defined as TN/(TN+FP). These correspond to the sensitivity on the positive and on the negative class respectively. For diagnostic tests it is crucial to know whether a test result is reliable. We capture this by two measures: (1) the negative predictive value NPV:=TN/(TN+FN) i.e. the precision on the negative class and (2) the positive predictive value PPV:=TP/(TP+FP). The PPV is often referred to as precision. We then compute a receiver operator characteristics curve (ROC). The ROC curve shows all possible thresholds on a numeric feature interpreted as classifier and shows the canonical FPR against TPR values. The area under this curve (AUROC) is 1.0 where prediction is optimal and the binary class is perfectly separated. The F-Measure is the harmonic mean of TPR and PPV i.e., (2*TPR*PPV)/(TPR+PPV).

4.4 Summary of Applied Methods

We applied all methods below as they are shipped with the WEKA system [11]. For libSVM (version 3.0, [6]) we use the wrapper provided by WEKA.
Support Vector Techniques. Support Vector Machines (SVM, [37, 39]) have become an integral part of statistical learning procedures. We apply Support Vector Classification (SVC) working on binary labels. The SVC model is a linear function in possibly high dimensional space – a hyperplane. Its placement is optimized to separate instances in two classes with a maximum margin. In this way so-called soft-margin SVMs or C-SVMs allow for misclassified instances during training. A penalty term C quantifies the weight for an instance. For an SVM, each instance is encoded as a vector of features corresponding to rows of F. Instances are compared via a *kernel* function. We use a linear kernel, which is simply a scalar product $< F_i, F_j >$, and we also apply a high dimensional kernel - the radial basis function (RBF Kernel, described in reference [6]).
Logistic Regression Models. Most existing risk stratifiers build upon logistic regression models. For k classes they model the posterior probability of each class via linear functions in the measured features F_i for a patient i (See [13] for a more detailed introduction). For $k = 2$ a set of linear functions with $l \in L = \{0, 1\}$

$$\beta_{1j} + \beta_{2j} * F_{ij} \quad = \quad \log \frac{P(l = 0|F_{ij})}{P(l = 1|F_{ij})} \, \forall j = 1...m \tag{1}$$

is fitted. For the commonly used risk stratifiers further feature selection and discretization procedures are applied, yet all of them boil down to a logistic regression model. In order to have a baseline comparator for our models, we computed AUCs for SAPSI scores for each ICU admission. Although the SAPSI would be considered by the medical community to have been superseded by later scores, it was not possible to use scores such as APACHE II, APACHE III or SAPS II because they are either proprietary or because some necessary markers are based on expert medical opinion and are not routinely captured in most patient data management systems.
Decision Trees. Decision trees are tree-like classifiers where each leaf represents a label $l \in L$. We use the grafted C4.5 variant shipped with WEKA termed *J48graft*.
Random Forests. Random Forests are ensemble classifiers which build several decision trees and use a majority voting strategy to arrive at a decision. Each tree is build from a subset of parameters (or instances in some formulations) yielding quite stable predictions.
Unbalanced Classes. A problem one often faces in data mining settings are unbalanced label assignments within the dataset. Models that focus on the maximization of correctly predicted

instances while minimizing false predictions tend towards prediction of the majority class. Accordingly it is crucial to reduce the weight of the larger class. LibSVM allows class weighting directly, while for other algorithms we use the WEKA *CostSensitiveClassifier*. It provides a wrapped cost function for arbitrary classifiers. We choose the inverse fraction of the training set class distribution as the weighting.

5 Results

We divide the patients into two sets designated the training and validation sets. The training set contains 797 patients and the validation set contains 749 patients. The validation set serves as an independent testing set. In the training set there were 661 survivors ($l = 0$) and 136 mortalities ($l = 1$). The validation set contained 631 patients labeled $l = 0$ and 118 labelled $l = 1$. Were majority class prediction applied to all patients, this would correspond to a precision of 82.9% in the training and 84.2% in the validation set.

As feature set F we extract a set of parameters that are known to be related to outcome in critical illness. Missing values are replaced by their column's median. Variables include the mean value for the parameter over the first 24 hours of the hospital admission; time series values for heart rate and blood pressure (the mean value for each 4 hour period in the first 48 hours of ICU admission); a mean equipotent to noradrenaline dosage of the inotropic agents noradrenaline, adrenaline, dopamine, vasopressin and phenylephrine during the first 24 hours of admissions; a score based on the International Classification of Diseases diagnoses assigned to each patient and a number of factors related to clinical history and basic patient characteristics such as age and weight. Concerning the time series data, each time point is treated as a separate feature. We choose to use the SAPS I [9] and SOFA [42] scores as baseline performance comparators for our own predictions. It should be pointed out that these scores have been superseeded by newer scores such as the SAPSII and APACHE III scores. We are unable however to retrieve these newer scores form the MIMIC-II database because, in contrast to SOFA and SAPSI, they require 'expert medical opinion' that is not routinely captured in clinical data management systems.

From a feature matrix F containing one patient's parameters per row we build a model M for each method. $M(F_i)$ for a patient i predicts the mortality during hospital stay. A standardized value heatmap of the matrix M is shown for all 37 input variables (see Section 4) in Figure 1. We note that no obvious patterns are discernable. Considering the heart rate timeline (2) confirms conventional medical wisdom that extreme heart rates are related to poor prognosis.

We apply a 10-fold cross validation (10-CV) on the training set as described in Section 4. Table 2 shows the area under curve values (AUC) and further statistics for decision trees (DT), random forests (RF), logistic regression (LR) and support vector classification with both linear kernel (SCLin) and RBF kernel (SCRBF) in a 10-fold CV. In Figure 3 we show the ROC curves for all classifiers with respect to mortality prediction as predicted class. For some classifiers the curves have a lower resolution due to non-continuous decision values. Next to the 10-CV, the independent validation set performance is reported in Table 3. We apply all algorithms in both *normal* and a *cost sensitive* (c.s.) version. We choose penalties of 6 (false negatives) and 1 (false positives), respectively. These values represent the class distribution (the survivors to mortalities ratio was approximately 6:1) within the training set.

[2] these scores cannot be be recomputed and are directly taken from the MIMICII database

Cost Sensitive	Method	TNR	NPV$^{\blacktriangle}$	TPR	PPV	F-Measure	AUROC
Yes	SCRBF	0.722	$0.946^{(1)}$	$0.801^{(1)}$	0.372	$0.508^{(2)}$	0.762
Yes	SCLin	0.749	$0.945^{(2)}$	$0.787^{(2)}$	0.392	$0.523^{(1)}$	0.768
Yes	LR	0.756	$0.928^{(3)}$	$0.713^{(3)}$	0.376	$0.492^{(3)}$	$0.81^{(1)}$
Yes	DT	0.893	0.877	0.39	0.427	0.408	0.644
Yes	RF	0.92	0.867	0.31	0.448	0.371	$0.809^{(2)}$
No	LR	0.927	0.868	0.316	$0.473^{(3)}$	0.379	0.801
No	DT	0.92	0.886	0.426	$0.523^{(2)}$	0.47	0.642
No	RF	$0.98^{(3)}$	0.858	0.213	$0.69^{(1)}$	0.326	$0.805^{(3)}$
No	SCLin	$0.98^{(2)}$	0.858	0.213	$0.69^{(1)}$	0.326	0.597
No	SCRBF	$1.0^{(1)}$	0.829	0.0	0.0	0.0	0.5
-	SAPSI [2]	-	-	-	-	-	0.694
-	SOFA [2]	-	-	-	-	-	0.648

Table 2: **Cross validation on training set.** Negative Predictive Value (NPV, precision in predicted negatives), Positive Predictive Value (PPV, precision in predicted positives) F-Measure and the area under curve (AUC). Cost sensitive variants of each algorithm (i.e. informed of the underlying class distribution) are included. The parameter settings are WEKA defaults. In each column, the three best scores are indicated by a rank in brackets.

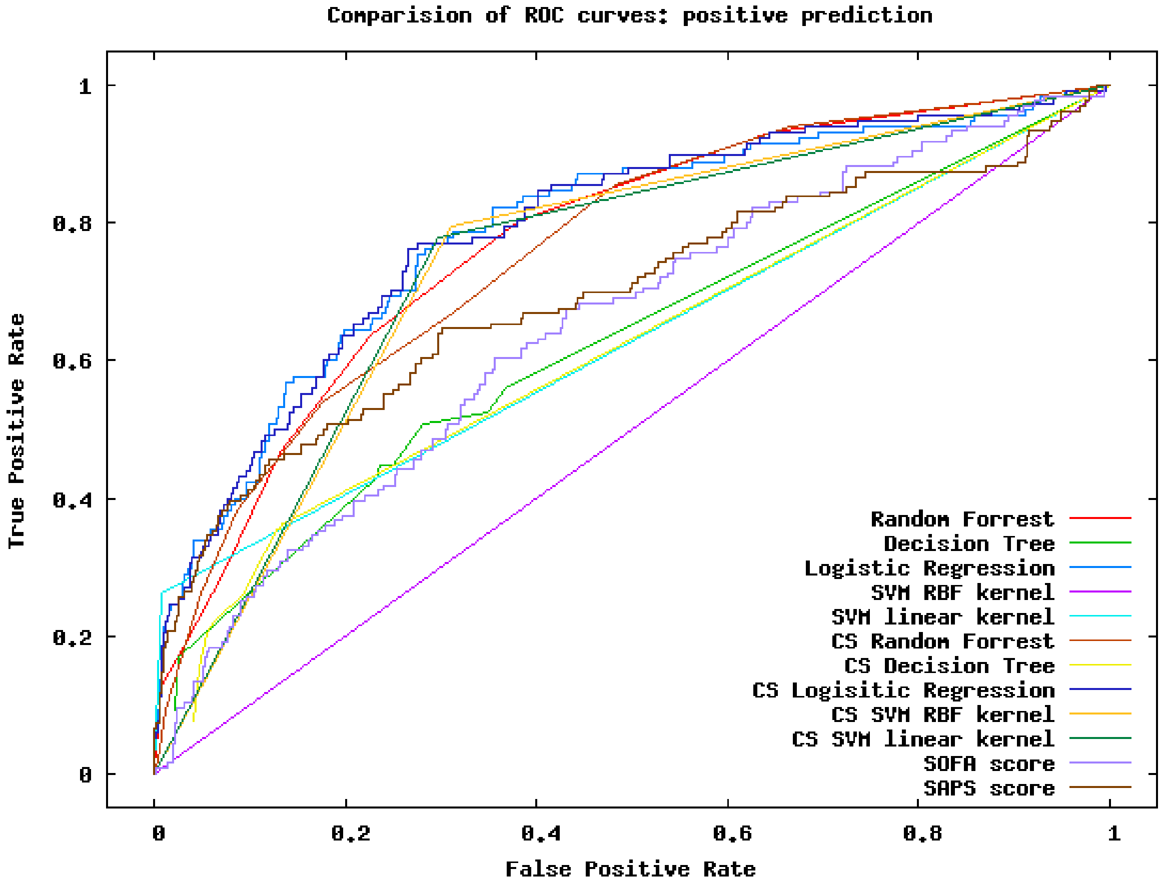
5.1 Unweighted Classes

In case of unweighted algorithms (that is where the algorithm was not informed of the class distribution) the support vector variants perform poorly regarding AUCs and F-Measures. They tend too strongly towards predict the majority class (reflected by high positive predictive values) and the resulting AUCs are close to 0.5. In fact, the SCRBF performance is no better than random both during testing and validation. This is not surprising because the support vector classifier is designed to maximize the amount of correctly classified instances rather than AUC. With respect to AUCs, the best performers in the unweighted algorithms were LR and random forrests, which had an AUCs of 0.80 and 0.81 respectively. This values are consistent with values reported in the literature for existing risk stratifiers.

5.2 Weighted Classes

For most of the classifiers it is possible to trade decreased false negatives (FN) against increased false positives (FP) with little impact on either true positives (TP) or true negatives (TN). By forcing unequal class weights we primarily observed an improvement in the rate of FNs. In the context of intensive care medicine this is highly desirable. The negative predictive value (NPV) is the amount of error when the test outcome is negative. In case of the validation set, the expected value, were all cases to be predicted as negative, would be 0.842 (the support vector classifier with an RBF kernel predicted all as negative and had an NPV of 0.842). The corresponds to a recall on the negative class (TNR) of 1.0. We observe that all methods except c.s. SCRBF, SCLin and LR have a relatively high type II error rate (FN) drastically reducing the TPR. This is however associated with a better positive predictive value (PPV). The best precision of 0.455 is achieved by c.s. RF at a recall of 0.381. The best recall is the c.s. SCRBF with 0.797 at a precision of 0.324. The cost sensitive LR achieves the best F-measure and AUROC, but its TPR is almost 10% below that of support vector machines with similar PPV. Notably, its NPV is ranked third. SAPSI and SOFA as instances of traditional risk stratifiers

Figure 3: ROC curves of all applied classifiers on the validation set. 'CS' denotes cost-sensitive versions of the algorithms using either class weighting or the WEKA cost-sensitive classifier wrapping procedure. Abbreviations: decision trees (DT), random forests (RF), logistic regression (LR) and support vector classification with both linear kernel (SCLin) and RBF kernel (SCRBF).



trained on a large set of patients show medium AUROC values.

6 Discussion

In this work we present a comparison of risk stratifiers for mortality prediction derived from automatically monitored parameters of patients in Intensive Care Units (ICUs). Specifically, we compare classic logistic regression models with other machine learning tools like support vector classification and random forests. The models are trained on a publicly available patient data set: the MIMICII dataset [36]. In contrast to the design of existing models we include time series parameters in our model (4 hourly heart rate and blood pressure measurements during the first 48 hours of the ICU admission). From the database we extract a subset of 1546 patients, 797 of which we use for training. The remainder serve as an independent validation set. Regarding mortality, the data set is unbalanced. Approximately 14% of patients are ”positive” (died in the hospital) and 86% were ”negative” (discharged from hospital alive and well). We show that measures to correct this imbalance profoundly affect the performance of all algorithms. In order to perform a fair comparison of all algorithms we apply cost-sensitive

Cost Sensitive	Method	TNR	NPV$^{\blacktriangle}$	TPR	PPV	F-Measure	AUROC
Yes	SCRBF	0.689	$0.948^{(1)}$	$0.797^{(1)}$	0.324	$0.461^{(3)}$	0.743
Yes	SCLin	0.704	$0.945^{(2)}$	$0.78^{(2)}$	0.33	$0.463^{(2)}$	0.742
Yes	LR	0.751	$0.931^{(3)}$	$0.703^{(3)}$	0.346	$0.464^{(1)}$	$0.791^{(1)}$
Yes	RF	0.914	0.888	0.381	0.455	0.415	0.762
No	LR	0.933	0.886	0.356	$0.5^{(3)}$	0.416	$0.789^{(2)}$
No	DT	0.764	0.881	0.449	0.262	0.331	0.625
Yes	DT	0.872	0.879	0.356	0.341	0.349	0.616
No	SCLin	$0.992^{(2)}$	0.878	0.263	$0.861^{(1)}$	0.403	0.627
No	RF	$0.968^{(3)}$	0.864	0.186	$0.524^{(2)}$	0.275	$0.772^{(3)}$
No	SCRBF	$1.0^{(1)}$	0.842	0.0	0.0	0.0	0.5
-	SAPSI ^{2}	-	-	-	-	-	0.684
-	SOFA ^{2}	-	-	-	-	-	0.640

Table 3: **Validation set results.** Statistics are described in Table 2. Note that the cost insensitive variant of the linear support vector classification offers both high negative and positive predictive value. Yet, only a quarter of all positives is detected (TPR). The table clearly shows the possibility to sacrifice negative for positive predictive value.

meta-classifiers, effectively simulating a 1:1 mortality distribution.

Both the logistic regression and support vector machine models perform well, yet no clear winner can be chosen. We emphase that the negative predictive values (NPV) of the best performing models seem comparable but have to be looked at very closely. For instance, in our data set, differences in NPV of 1.7% between LR and SCRBF (cost sensitive) correspond to 24 vs. 35 patients incorrectly classified negative. On the other hand the positive predictive values (PPV: proportion of correct positive predictions) are higher with SVMs that are not aware of class distribution. These higher PPVs are however at the cost of lower total positive rates (fewer positive cases in total).

A predictive model with a high NPV only rarely misclassifies a patient with a good prognosis as having a poor prognosis. Misclassification of a patient with a good prognosis could have catastrophic consequences in an ICU, were for instance a withdrawal of care decision to be made based on the predictive model. From a critical care standpoint, a high NPV is an indispensable characteristic of any predictive model. The cost sensitive linear support vector classificator detects 70% of all negatives at an NPV of 94.5% i.e., 6 out of 100 patients will be falsely classified as negatives. This improvement comes at the high price of reduced PPV. Ranking by AUROC suggest Random Forests to be among the best methods, yet the NPV of 86.4% corresponds to 96 false negatives (in comparison to 24 with the best method).

As discussed in the introduction, a focus of the criticism of risk prediction models currently available is that they are poorly calibrated – their predictions do not reflect the true probability of death or survival on an individual patient basis. Our results emphasise that the focus should not purely be on calibration - there are clinically important parameters such as negative predicitive value that such also be taken into account. Out data demonstrate that when considering any model, it is vital that the way the model will be used in practive is taken into account. A model capable of predicting patients with good prognosis at high precision and recall may in fact be more useful than one predicting only a fraction of the patients with poor prognosis with 100% certainty. Thus, models have to be chosen with great care and after taking into account their future clinical use.

A focus of future work should be risk prediction models that are updated over time in an automated manner [44]. Previously trained models that continually evaluate novel patient data are likely capable of very specific predictions. Training models specific for specific class predictions and specific patient groups, and employing the latest available patient data, will be extremely helpful for intensive care medicine. Furthermore continuous re-evaluation and re-training is extremely important. It has been shown that performance falls with time: modern predictive models will have to cope with rapid changes due to improved therapies, patient characteristics and varying patient groups.

In this work, we have shown that model choice and even meta-critera such as class specific costs make a significant difference regarding predictive capablities. The implementation of flexible, specific models in ICUs is a perhaps a distant goal but is definitely worth pursuing. Given appropriate validation sets for model classes on a ward- , hospital- and even country-specific scale, patient specific models rooted in machine learning techniques are feasible objectives.

References

[1] D. H. Beck, G. B. Smith, J. V. Pappachan, and B. Millar. External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: a multicentre study. *Intensive Care Med*, 29(2):249–256, Feb 2003.

[2] C. D. Bufalo, A. Morelli, L. Bassein, L. Fasano, C. C. Quarta, A. M. Pacilli, and G. Gunella. Severity scores in respiratory intensive care: APACHE II predicted mortality better than SAPS II. *Respir Care*, 40(10):1042–1047, Oct 1995.

[3] M. Capuzzo, V. Valpondi, A. Sgarbi, S. Bortolazzi, V. Pavoni, G. Gilli, G. Candini, G. Gritti, and R. Alvisi. Validation of severity scoring systems SAPS II and APACHE II in a single-center population. *Intensive Care Med*, 26(12):1779–1785, Dec 2000.

[4] X. Castella, A. Artigas, J. Bion, and A. Kari. A comparison of severity of illness scoring systems for intensive care unit patients: results of a multicenter, multinational study. The European/North American Severity Study Group. *Crit Care Med*, 23(8):1327–1335, Aug 1995.

[5] X. Castella, J. Gilabert, F. Torner, and C. Torres. Mortality prediction models in intensive care: acute physiology and chronic health evaluation II and mortality prediction model compared. *Crit Care Med*, 19(2):191–197, Feb 1991.

[6] C. Chang and C. Lin. LIBSVM: a library for support vector machines., 2001.

[7] N. R. Cook. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, 115(7):928–935, Feb 2007.

[8] L. Esserman, J. Belkora, and L. Lenert. Potentially ineffective care. A new outcome to assess the limits of critical care. *Jama*, 274(19):1544–51, 1995.

[9] J. R. L. Gall, P. Loirat, A. Alperovitch, P. Glaser, C. Granthil, D. Mathieu, P. Mercier, R. Thomas, and D. Villers. A simplified acute physiology score for ICU patients. *Crit Care Med*, 12(11):975–977, Nov 1984.

[10] L. Gortzis, F. Sakellaropoulos, I. Ilias, K. Stamoulis, and I. Dimopoulou. Predicting ICU survival: a meta-level approach. *BMC Health Serv Res*, 8:157, 2008.

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[12] E. Hanisch, R. Brause, J. Paetz, and B. Arlt. Review of a large clinical series: Predicting death for patients with abdominal septic shock. *J Intensive Care Med*, 26(1):27–33, 2011.

[13] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning.* Springer, 2001.

[14] J. Hunt and A. Meyer. Predicting survival in the intensive care unit. *Curr Probl Surg*, 34(7):527–99, 1997.

[15] F. Jaimes, J. Farbiarz, D. Alvarez, and C. Martinez. Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room. *Crit Care*, 9(2):R150–56, 2005.

[16] R. Jayes, J. Zimmerman, D. Wagner, E. Draper, and W. Knaus. Do-not-resuscitate orders in intensive care units. Current practices and recent changes. *Jama*, 270(18):2213–7, 1993.

[17] X. Jia, A. Malhotra, M. Saeed, R. Mark, and D. Talmor. Risk factors for ARDS in patients receiving mechanical ventilation for ¿ 48 h. *Chest*, 133(4):853–61, 2008.

[18] S. Katsaragakis, K. Papadimitropoulos, P. Antonakis, S. Strergiopoulos, M. M. Konstadoulakis, and G. Androulakis. Comparison of Acute Physiology and Chronic Health Evaluation II (APACHE II) and Simplified Acute Physiology Score II (SAPS II) scoring systems in a single Greek intensive care unit. *Crit Care Med*, 28(2):426–432, Feb 2000.

[19] W. Knaus, D. Wagner, E. Draper, J. Zimmerman, M. Bergner, P. Bastos, C. Sirio, D. Murphy, T. Lotring, A. Damiano, and et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest*, 100(6):1619–36, 1991.

[20] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman. APACHE II: a severity of disease classification system. *Crit Care Med*, 13(10):818–829, Oct 1985.

[21] W. A. Knaus, J. E. Zimmerman, D. P. Wagner, E. A. Draper, and D. E. Lawrence. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. *Crit Care Med*, 9(8):591–597, Aug 1981.

[22] J. Le Gall, S. Lemeshow, and F. Saulnier. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *Jama*, 270(24):2957–63, 1993.

[23] S. Lemeshow, D. Teres, J. Klar, J. Avrunin, S. Gehlbach, and J. Rapoport. Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. *Jama*, 270(20):2478–86, 1993.

[24] S. Lin, C. Lee, Y. Lu, and L. Hsu. A comparison of MICU survival prediction using the logistic regression model and artificial neural network model. *J Nurs Res*, 14(4):306–14, 2006.

[25] Y. Liu, L.-Q. Wei, G.-Q. Li, F.-Y. Lv, H. Wang, Y.-H. Zhang, and W.-L. Cao. A decision-tree model for predicting extubation outcome in elderly patients after a successful spontaneous breathing trial. *Anesth Analg*, 111(5):1211–1218, Nov 2010.

[26] B. M. Livingston, F. N. MacKirdy, J. C. Howie, R. Jones, and J. D. Norrie. Assessment of the performance of five intensive care scoring models within a large Scottish database. *Crit Care Med*, 28(6):1820–1827, Jun 2000.

[27] O. Luaces, F. Taboada, G. Albaiceta, L. Dominguez, P. Enriquez, and A. Bahamonde. Predicting the probability of survival in intensive care unit patients from a small number of variables and training examples. *Artif Intell Med*, 45(1):63–76, 2009.

[28] R. Markgraf, G. Deutschinoff, L. Pientka, and T. Scholten. Comparison of acute physiology and chronic health evaluations II and III and simplified acute physiology score II: a prospective cohort study evaluating these methods to predict outcome in a German interdisciplinary intensive care unit. *Crit Care Med*, 28(1):26–33, Jan 2000.

[29] R. Moreno, G. Apolone, and D. R. Miranda. Evaluation of the uniformity of fit of general outcome prediction models. *Intensive Care Med*, 24(1):40–47, Jan 1998.

[30] A. Nimgaonkar, D. Karnad, S. Sudarshan, L. Ohno-Machado, and I. Kohane. Prediction of mortality in an Indian intensive care unit. Comparison between APACHE II and artificial neural networks. *Intensive Care Med*, 30(2):248–53, 2004.

[31] S. Nouira, M. Belghith, S. Elatrous, M. Jaafoura, M. Ellouzi, R. Boujdaria, M. Gahbiche, S. Bouchoucha, and F. Abroug. Predictive value of severity scoring systems: comparison of four models in Tunisian adult intensive care units. *Crit Care Med*, 26(5):852–859, May 1998.

[32] P. A. Patel and B. J. Grant. Application of mortality prediction systems to individual intensive care units. *Intensive Care Med*, 25(9):977–982, Sep 1999.

[33] A. Randolph, G. Guyatt, and J. Carlet. Understanding articles comparing outcomes among intensive care units to rate quality of care. Evidence Based Medicine in Critical Care Group. *Crit Care Med*, 26(4):773–81, 1998.

[34] K. M. Rowan, J. H. Kerr, E. Major, K. McPherson, A. Short, and M. P. Vessey. Intensive Care Society's Acute Physiology and Chronic Health Evaluation (APACHE II) study in Britain and Ireland: a prospective, multicenter, cohort study comparing two methods for predicting outcome for adult intensive care patients. *Crit Care Med*, 22(9):1392–1401, Sep 1994.

[35] M. Saeed. *Temporal pattern recognition in multiparameter ICU data.* PhD thesis, M.I.o.T., Department of Electrical Engineering and Computer Science, http://dspace.mit.edu/handle/1721.1/40507, 2007.

[36] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark. Multiparameter Intelligent Monitoring in Intensive Care II: A public-access intensive care unit database. *Crit Care Med*, 39(5):952–960, May 2011.

[37] A. J. Smola, B. Schlkopf, and K. R. Mller. The connection between regularization operators and support vector kernels. *Neural Netw*, 11(4):637–649, Jun 1998.

[38] I. K. Tan. APACHE II and SAPS II are poorly calibrated in a Hong Kong intensive care unit. *Ann Acad Med Singapore*, 27(3):318–322, May 1998.

[39] V. N. Vapnik. An overview of statistical learning theory. *IEEE Trans Neural Netw*, 10(5):988–999, 1999.

[40] M. J. Vassar, F. R. Lewis, J. A. Chambers, R. J. Mullins, P. E. O'Brien, J. A. Weigelt, M. T. Hoang, and J. W. Holcroft. Prediction of outcome in intensive care unit trauma patients: a multicenter study of Acute Physiology and Chronic Health Evaluation (APACHE), Trauma and Injury Severity Score (TRISS), and a 24-hour intensive care unit (ICU) point system. *J Trauma*, 47(2):324–329, Aug 1999.

[41] T. Verplancke, S. V. Looy, K. Steurbaut, D. Benoit, F. D. Turck, G. D. Moor, and J. Decruyenaere. A novel time series analysis approach for prediction of dialysis in critically ill patients using echo-state networks. *BMC Med Inform Decis Mak*, 10:4, 2010.

[42] J. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonca, H. Bruining, C. Reinhart, P. Suter, and L. Thijs. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22(7):707–710, 1996.

[43] P. Wilairatana, N. S. Noan, S. Chinprasatsak, K. Prodeengam, D. Kityaporn, and S. Looareesuwan. Scoring systems for predicting outcomes of critically ill patients in northeastern Thailand. *Southeast Asian J Trop Med Public Health*, 26(1):66–72, Mar 1995.

[44] S. L. Zeger, R. Irizarry, and R. D. Peng. On time series analysis of public health and biomedical data. *Annu Rev Public Health*, 27:57–79, 2006.