IOP Conference Series: Materials Science and Engineering

PAPER • OPEN ACCESS

Mortality prediction in critically ill patients using machine learning score

To cite this article: F Dzaharudin *et al* 2020 *IOP Conf. Ser.: Mater. Sci. Eng.* **788** 012029

Mortality prediction in critically ill patients using machine learning score

F Dzaharudin[1, *], A M Ralib[2], U K Jamaludin[1], M B M Nor[2], A Tumian[3], L C Har[4] and T C Ceng[5]

[1] Faculty of Mechanical Engineering, Universiti Malaysia Pahang, 26600 Pekan, Pahang, Malaysia
[2] Kulliyyah of Medicine, International Islamic University Malaysia, Kuantan, 25200, Malaysia.
[3] Kulliyyah of Information and Communication Technology, International Islamic University Malaysia, Gombak, 53100, Malaysia.
[4] Department of Anaesthesiology, Hospital Pulau Pinang, 10990, Pulau Pinang, Malaysia
[5] Department of Anaesthesiology and Intensive Care, Sultanah Aminah Hospital, 81300, Johor Bahru, Malaysia

*Corresponding author: fatimahd@ump.edu.my

Abstract. Scoring tools are often used to predict patient severity of illness and mortality in intensive care units (ICU). Accurate prediction is important in the clinical setting to ensure efficient management of limited resources. However, studies have shown that the scoring tools currently in use are limited in predictive value. The aim of this study is to develop a machine learning (ML) based algorithm to improve the prediction of patient mortality for Malaysian ICU and evaluate the algorithm to determine whether it improves mortality prediction relative to the Simplified Acute Physiology Score (SAPS II) and Sequential Organ Failure Assessment Score (SOFA) scores. Various types of classification algorithms in machine learning were investigated using common clinical variables extracted from patient records obtained from four major ICUs in Malaysia to predict mortality and assign patient mortality risk scores. The algorithm was validated with data obtained from a retrospective study on ICU patients in Malaysia. The performance was then assessed relative to prediction based on the SAPS II and SOFA scores by comparing the prediction accuracy, area under the curve (AUC) and sensitivity. It was found that the Decision Tree with SMOTE 500% with the inclusion of both SAPS II and SOFA score in the dataset could provide the highest confidence in categorizing patients into two outcomes: death and survival with a mean AUC of 0.9534 and a mean sensitivity 88.91%. The proposed ML score were found to have higher predictive power compared with ICU severity scores; SOFA and SAPS II.

Keywords. Machine learning, Mortality prediction, Severity, SAPS II score, SOFA score;

1. Introduction

Scoring systems are often used in the Intensive Care Unit (ICU) to assess the severity of the disease, compare ICU performances, in research and to predict mortality [1-3]. Prediction of patient mortality and severity is crucial for clinicians to make sound medical decisions for patient and resource management; which includes treatment, prevention and efficient allocation of limited resources. This is a necessity in busy ICUs since medical resources such as limited number of doctors, nurses and facilities, may not be sufficient for all the patients to be instantaneously attended to. A scoring model can stratify acutely ill patients who would more urgently require such resources and potentially benefit from it thus improving allocation of resources, clinical decision making leading to a better quality of care in the ICU.

Numerous scoring models have been developed to assess and characterize the severity of illness of critically ill patients. The scores generally belong to one of two categories [1-4]: (1) scores that aim to quantify the level of organ dysfunction daily during ICU stay, for example the Sequential Organ Failure Assessment (SOFA) [5] (2) score that aims to predict mortality based on parameters upon ICU admission or during the first 24 hours of ICU stay, for example, the Simplified Acute Physiology Score (SAPS II) [6].

The Sequential Organ Failure Assessment (SOFA) score was developed to describe changes in organ function and sequence of complications in the critically ill. The score is based on 6 variables; respiration, coagulation, liver, cardiovascular, renal and central nervous system (Glasgow Coma Score). A score of 0-4 is assigned for the six organ systems which describes the worst values for every 24-hour period in the ICU. Although the score was not designed to predict mortality, an association between increasing initial organ-specific SOFA scores and mortality has been suggested since mortality rate is related to the degree of organ dysfunction [4, 5]

The Simplified Acute Physiology Score (SAPS II) was developed and validated using data from 137 ICUs from 12 countries in a European and North American cohort (n = 13,152) [6]. The mortality risk is estimated based on the sum of the score. The score includes 17 variables consisting of 12 physiological variables, type of admission (scheduled/unscheduled surgical or medical), age and 3 underlying disease variables hematologic malignancy, metastatic cancer and acquired immunodeficiency syndrome).

To this day, scoring tools such as SAPS II and SOFA scores remain widely used in clinical practice and have been known to discriminate survivors and non-survivors well [7]. Inevitably, these models reflect the medical culture and the population characteristics which they originate from. The SAPS II severity score for mortality estimation, was developed from a large sample of ICU patients from North America and Europe [5, 6]. Although the methods have been showed to adapt well and result in good discrimination when applied to new populations from other institutions, external validation studies performed in various countries have shown that even recent versions of critical warning scoring tools are not adequately calibrated and have large variability in accuracy across various diseases and populations [8-13]

These scoring models, were not designed to be sensitive to the underlying physiology of individual patients and does not factor in variations in patient information trends [5]. This is problematic when applying predictions to a larger scale due to the complex and heterogeneous mixture of patients and diagnoses thus leading to locally-customised variants of these scores to improve prediction. As a result, the performance of SAPS II, for example has been found to be affected by case-mix [3] [14] and national differences. This has led to the development of various versions of SAPS. For example, there has been research studies to specifically tailor to France, to Southern Europe and Mediterranean countries and to Central and Western Europe (see [15-17]) and for very elderly ICU patients [18].

Furthermore, these scores do not account for changes in diagnostic, therapeutic and prognostic techniques. The SOFA and SAPS II score for example were developed in 1996 [5] and 1993 [6] respectively. There has been much progress in diagnostic, therapeutic and prognostic techniques since. Thus, to ensure accuracy of the scoring models in today's ICU, there's is a need for the current scoring models to be continuously updated as new diagnostic, therapeutic and prognostic techniques are developed [1].

In this study, we aim to develop a novel ML score for risk stratification of critically ill patients presented to the Malaysian ICU for mortality prediction. Instead of developing locally-customized variants of the common scoring tools, we will leverage the use of multidimensional analysis which includes various parameters, such as physiological measurements, admission types etc., as well as SAPS II [6] and SOFA [1] scores, all data which are readily available in the Malaysian ICU database to improve accuracy of prediction. We will also show that including SAPS II and SOFA scores will improve ML prediction for patient mortality.

2. Methodology

2.1. Study design and patient recruitment eligibility

A prospective, nonrandomised, observational cohort study was implemented consisting of a database of 28,790 critically ill patients above the age of 18 years old, admitted to four hospitals in Malaysia; Penang General Hospital, Kuala Lumpur Hospital, Sultanah Aminah Hospital and Tengku Ampuan Afzan Hospital between January 2010 to December 2014. These hospitals are the main public hospitals in their respective states and among the covering the south, north, west and east coast of Malaysia. The baseline characteristics is given in Table 1. Inclusion criteria for this dataset used in this study were patients with more than one observation for each clinical variable outlined in table 1 and the existence of values for SOFA and SAPS II scores. Patients without these scores were removed from the dataset to allow comparison between the performance in mortality prediction between SOFA, SAPS II and the algorithm developed in this study.

Following this patient inclusion process, a total of 25,524 patients were selected in this study, with median SAPS II 49 (IQR: 35 - 63), median SOFA 9 (IQR: 6 - 12). A total of 4,304/21,220 (20.29%) patients died in the intensive care units. The study was registered with the National Medical Research Registration (NMRR 14-1938-23183) and ethics approval was obtained from the Malaysian Research Ethics Committee (MREC) with a waiver of patient consent.

Since the primary goal of this study is to predict mortality after ICU admission, the ICU discharge status was grouped into two bins; either 'Alive' or 'Death'. Patients who were labelled as 'Discharged with grave prognosis' and 'Transferred to another hospital' were labelled as 'Alive'. By grouping these labels, we could distinctly predict and score the desired outcome. Patients were followed up until discharged or in-hospital death. Model construction used data from 17,846 patients whereas model validation used data from 7,649 patients. The partitioning was also done with a built-in method in Azure Machine Learning which randomized the patient based on the patient outcome.

3. Results and discussion

The gold standard used in this study was the in-ICU mortality. The definition of this gold standard classified 4,304 patients as having in-ICU deaths and 21,220 patients as survivors which results in a prevalence of 16.86% in-ICU mortality as shown in table 1. Here, we have an imbalanced dataset which consists of a group with a majority of normal samples and a minority of samples with abnormal outcome (death). Imbalanced data may cause standard classifiers such as logistic regression, Support Vector Machine and decision trees to provide suboptimal classification results where the majority class has good coverage whereas the minority class are distorted. For this work, Synthetic Minority Oversampling Technique (SMOTE) [19] was implemented to cope with imbalanced data classification. SMOTE is an oversampling method which works by increasing the number of the minority class through random data replication.

To avoid over-fitting while increasing the minority class, the SMOTE built-in module in Azure Machine Learning uses imputation to statistically sample the rows and impute the minority label within the existing feature space. By deriving new instance values from interpolation instead of extrapolation will ensure relevance to the underlying data set.

Table 1. Baseline characteristics of 25, 524 patients included in the study. Data are presented as mean ± standard deviation or number (%).

Variable	No mortality within ICU (n=21,220)	Mortality within ICU (n=4,304)
Age (years)	47.75 ± 17.48	52.21 ± 16.54
Male gender	12,603 (59.39%)	2784 (35.29%)
Ethnicity		
Malay	11,237 (52.95)	2,300 (53.44)
Chinese	4,724 (22.26)	960 (22.30)
Indian	2,882 (13.58)	600 (13.94)
Others	2,377 (11.20)	444 (10.31)
ICU length of stay (hour)	125.57 ± 223.33	160.49 ±255.78
Hospital length of stay (hour)	486.51 ± 648.98	256.49 ± 388.59
Patient Category		
Non-Operative	12,510 (58.95)	3,088 (71.74)
Emergency Operative	6,046 (28.49)	1,091 (25.35)
Elective Operative	2,637 (12.43)	115 (2.67)
Others	27 (0.13)	10 (0.23)
Location before ICU admission		
Ward	7461 (35.16)	2034 (47.26)
Operation Theatre	7352 (34.65)	908 (21.10)
Emergency Department	4603 (21.69)	931 (21.63)
Others	1803 (8.5%)	431 (10.02)
Main organ failures		
Cardiovascular	4877 (22.98)	2062 (47.91)
Neurological	2285 (10.77)	500 (11.62)
Haematological	964 (4.54)	100 (2.32)
Hepatic	165 (0.78)	34 (0.79)
Respiratory	3606 (16.99)	1190 (27.65)
Renal	1538 (7.25)	233 (5.41)
No organ failure	7785 (36.69)	185 (4.30)
Number of organ failures		
None	7906 (37.26)	195 (4.53)
One	7062 (33.28)	783 (18.19)
Two	4382 (20.65)	1521 (35.34)
Three	1551 (7.31)	1234 (28.67)
More than three	319 (1.50)	571 (13.27)
Severe sepsis within 24 hours of ICU admission	5699 (51.36)	2599 (45.68)
Acute respiratory distress syndrome (ARDS) within 24 hours of ICU admission	1401 (12.63)	958 (16.84)
Acute kidney injury (AKI) within 24 hours of ICU admission	3997 (36.02)	2133 (37.49)
SAPS II Score	34.00 ± 15.74	57.27 ± 17.71
SOFA Score	6.00 ± 3.7	11.16 ± 3.77

To evaluate the effectiveness of the SMOTE sampling technique when dealing with the problem of an imbalanced dataset, seven prediction models were built without and with the SMOTE sampling technique (100%, 300% and 500%). The prediction performance of different prediction models was measured using two evaluation metrics: area under curve (Figure 1) and sensitivity (Figure 2). These measures are often used to assess the performance of models in medical applications (see [20].

Furthermore, the choice of AUC as performance indicator is justified due to the imbalanced dataset which renders performance criteria such as accuracy inadequate to assess models' the performance as it tends to give advantage to models that output the class with highest frequency [21]. Sensitivity was also chosen as a performance indicator following Ong et al. [20].

The seven different machine learning models were trained using Support Vector Machine (SVN), Neural Network (NN), Logistic Regression (LR), Locally-Deep Support Vector Machine (LDSVN), Decision Forest (DF), Boosted Decision Tree (DT) and Boosted Decision Jungle (DJ). To gain insight on how SOFA and SAPS II score may contribute in improving prediction performance, the analysis performed for four cases:

- Case 1: Dataset without inclusion of SOFA and SAPS II score (blue boxplots in figure 1 and figure 2)
- Case 2: Inclusion of SOFA score in the dataset (red boxplots in figure 1 and figure 2)
- Case 3: Inclusion of SAPS II score in the dataset (orange boxplots in 1 and figure 2)
- Case 4: Inclusion of SOFA and SAPS II score in the dataset (green boxplots in figure 1 and figure 2)

Results show that the AUC (Figure 1) and sensitivity (figure 2) for all models using SMOTE resulted in significant improvement over the training results without SMOTE. In general, Boosted Decision Tree resulted in the best performance for all cases, with and without SMOTE for AUC and sensitivity (see columns DT in figure. 1 and 2 respectively). Models such as Decision Forest, Boosted Decision Trees and Boosted Decision Jungle are mostly on the higher side indicating superior performance than other models. It can also be seen that increasing the percentage of synthetic examples improves sensitivity and generally, AUC. For example, in Case 4, the DF model achieves a mean AUC of 0.9467 and a mean sensitivity of 73.58% using the sampled dataset with 100% synthetic examples compared to 0.9492 (mean AUC) and 83.25% (mean sensitivity) using the sampled dataset with 300% synthetic examples. Increasing the percentage of synthetic examples to 500% improves the AUC of DF to achieve 0.9534 (mean AUC) and 88.91% (mean sensitivity). For Case 1, DT showed the best improvement using SMOTE 100% achieving 0.9568 (mean AUC) compared to 0.9543 and 0.9523 using 300% and 500% synthetic examples respectively. However, in terms of mean sensitivity, DT (Case 1) achieved 91.61% using the sampled dataset with 500% synthetic examples compared to 87.74% (SMOTE 300%) and 82.74% (SMOTE 100%). This shows that the performance of each model can differ from one metric to another.

To assess the stability of the model prediction across the different cases, the best performance for each model was averaged after applying the ten-fold cross-validation on the training dataset. Here, the dataset is split into ten mutually exclusive and exhaustive blocks which are approximately equal in size. Nine blocks are then trained with each algorithm (the training set) and used to predict the outcome in the remaining block (the validation set) prior to the calculation of the mean squared error between the predicted and observed outcomes. To ensure that no patient appears in both the training and validation sets for each iteration and that for every patient observation to serve exactly once in the validation set and included in the training set at other times, this procedure was repeated 10 times using a different block as the validation set every round. This is to mitigate overfitting; an occurrence in which the algorithm is overly tailored to the available data at the expense of performance of external data, which is more likely to occur when training and validation sets intersect. The performance is then measured for each iteration and aggregated over all 10 iterations.

The Boosted Decision Tree model using SMOTE 500% for Case 4 was found to have the best mean in terms of AUC (0.9663 ± 0.022) and sensitivity (91.61% ± 0.5655%). This is also illustrated by the highest median shown by the boxplots in red in figure 1 and 2. It is noted that all the models with SMOTE achieved a more balanced sensitivity.

A large performance variance between the models may be observed for the models without SMOTE sample, which may be explained by the data imbalance between the two data classes (16.86%, 83.14%). Model variance is lower for the cases where SMOTE sampling was implemented due to increase in balance between the two data classes. In addition, the performance of each model can differ from one metric to another. Here, Neural Network without using SMOTE sampling showed the worst AUC performance for all 4 cases (second column in figure 1) with mean AUC of 0.7987 (Case 1), 0.8673 (Case 2), 0.8240 (Case 3) and 0.8337 (Case 4). However, for sensitivity, the Support Vector Machine without SMOTE showed the worst performance (first column in figure 2 with mean sensitivity of 24.05% (Case 1), 27.33% (Case 2), 36.89% (Case 3) and 37.06% (Case 4). These performance measures provided the lower bound of performance and hence, the level of difficulty of the prediction problem that we attempting to deal with. This is due to the algorithms being sensitive to data sampling and the number of neighbours. By including these methods in the analysis provides an indication of the lower bound of performance and hence the level of difficulty of the prediction problem in hand.

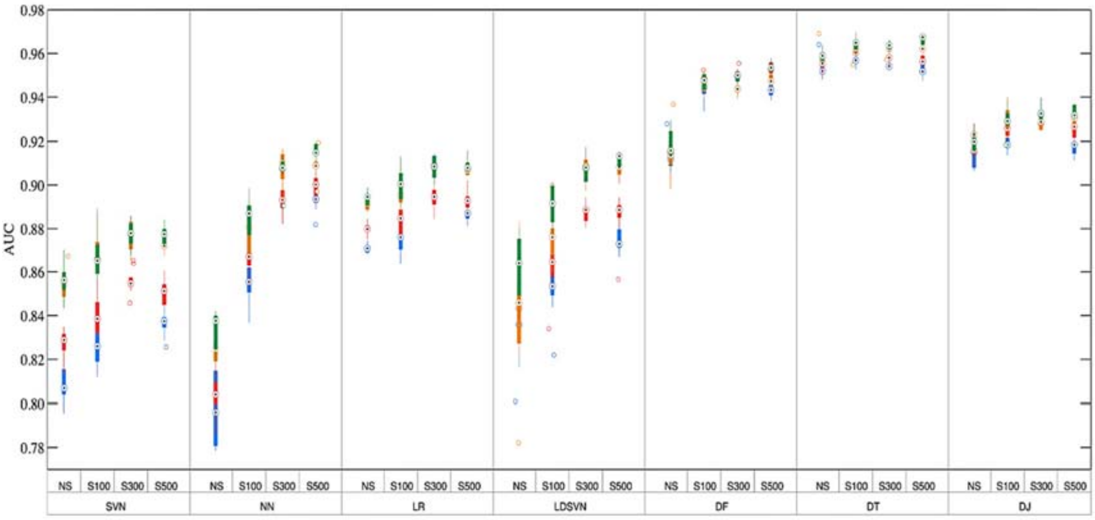

Figure. 1 AUC for Cross validation of n=10 folds without SMOTE (NS), SMOTE 100% (S100), SMOTE 300% (S300) and SMOTE 500% (S500) for (Blue) Case 1: dataset without SOFA and SAPS II score (Green) Case 2: dataset with SOFA score (Orange) Case 3: dataset with SAPS II score and (Red) Case 4: dataset with SOFA and SAPS II score. The seven machine learning models were trained using Support Vector Machine (SVN), Neural Network (NN), Logistic Regression (LR), Locally-Deep Support Vector Machine (LDSVN), Decision Forest (DF), Boosted Decision Tree (DT) and Boosted Decision Jungle (DJ).

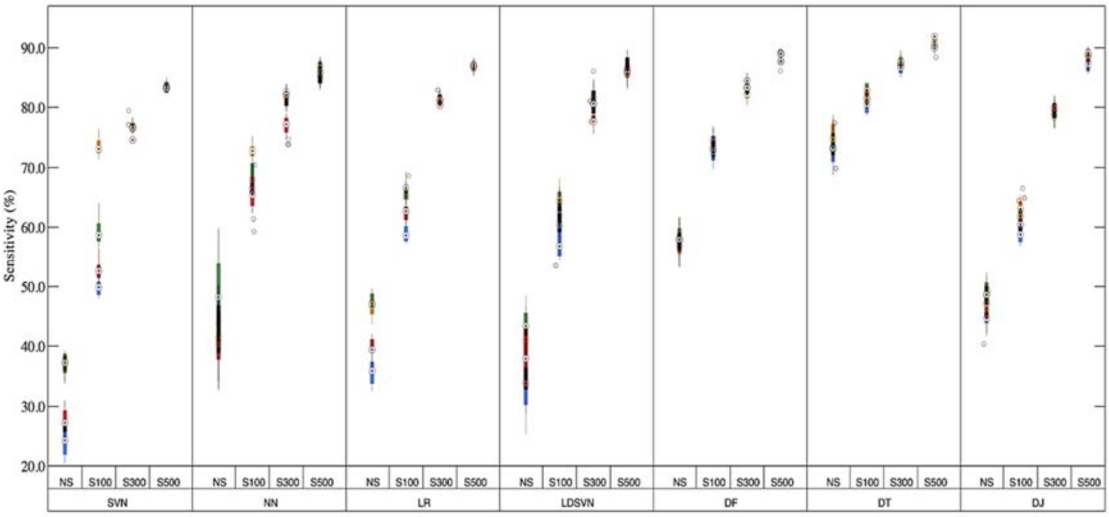

Figure. 2 Sensitivity for Cross validation of n=10 folds without SMOTE (NS), SMOTE 100% (S100), SMOTE 300% (S300) and SMOTE 500% (S500) for (Blue) Case 1: dataset without SOFA and SAPS II score (Green) Case 2: dataset with SOFA score (Orange) Case 3: dataset with SAPS II score and (Red) Case 4: dataset with SOFA and SAPS II score. The seven machine learning models were trained using Support Vector Machine (SVN), Neural Network (NN), Logistic Regression (LR), Locally-Deep Support Vector Machine (LDSVN), Decision Forest (DF), Boosted Decision Tree (DT) and Boosted Decision Jungle (DJ).

The numerical representation in terms of the best performance average are shown in tables 2 and 3. For each case, the highest value was highlighted in bold font. The Boosted Decision Tree model using SMOTE 500\% for Case 4 was found to have the best mean in terms of AUC (0.9663 ± 0.022) and sensitivity (91.61% ± 0.5655%), as well as the highest median shown by the red boxplots in the DT column in both figure 1 and 2. All the models with SMOTE achieved a more balanced sensitivity.

By comparing the performance average of the best models for the different cases highlighted in bold for AUC, (Table 2) and Sensitivity, (Table 3), it can be observed that by including SAPS II scores (Case 3) into the dataset, the prediction is improved 0.46 to 0.56 and 0.63% to 1.18% for AUC and Sensitivity respectively. This can also be seen in figure 1 and 2 where the lower and upper bounds of the orange boxplots (inclusion of SAPS II score) are higher for both AUC and sensitivity compared to when SAPS II score is excluded (see boxplots in blue and green in figure 1 and 2). This suggests that SAPS II score has a greater influence on the overall improvement in prediction compared to SOFA score. This may be due to the fact that SOFA scores were developed for sequentially assessment of the severity of organ failure during ICU stay and, though it may be used to predict mortality in various clinical conditions, it was not designed for that purpose [1].

The best prediction, however, is obtained by including both SOFA and SAPS II scores, Case 4, over other cases with prediction improvement of 0.39 to 0.95 average AUC and 0.64% to 1.82% average sensitivity. It can be concluded that the prediction is most accurate when both SOFA and SAPS II scores are included in the dataset (Case 4) as it has the highest performance.

Table 2. Best performance average summary for each model in terms of AUC. For each case, the highest value is highlighted in bold font. The values that were underlined, in italic and normal fonts were achieved with SMOTE 100%, SMOTE 300% and SMOTE 500% respectively.

Classification	Case 1	Case 2	Case 3	Case 4
Support vector machine	0.8774 ± 0.0062	0.8559 ± 0.0057	0.8766 ± 0.0075	0.8774 ±0.0062
Neural network	0.9070 ± 0.0064	0.9000 ± 0.0060	0.9092 ± 0.0057	0.9153 ± 0.0028
Logistic regression	0.9082 ± 0.0050	0.8930 ± 0.0046	0.9078 ± 0.0049	0.9084 ± 0.0034
Locally-deep SVN	0.9074 ± 0.0052	0.8853 ± 0.0109	0.9080 ± 0.0043	0.9115 ± 0.0034
Decision forest	0.9492 ± 0.0024	0.9526 ± 0.0032	0.9478 ± 0.0031	0.9534 ± 0.0030
Boosted decision tree	**0.9568** ± **0.0024**	**0.9578** ± **0.0016**	**0.9624** ± **0.0024**	**0.9663** ± **0.0022**
Boosted decision jungle	0.9322 ±0.0039	0.9280 ± 0.0015	0.9305 ± 0.0040	0.9329 ± 0.0034

Table 3. Best performance average summary for each model in terms of Sensitivity (%). For each case, the highest value for each case are highlighted in bold font and were achieved using SMOTE 500%.

Classification	Case 1	Case 2	Case 3	Case 4
Support vector machine	83.36 ± 1.0517	83.67 ± 0.6948	83.45 ± 0.0075	83.27 ± 0.6357
Neural network	85.51 ± 1.7860	84.83 ± 0.9879	86.03 ± 0.0057	86.10 ± 1.8253
Logistic regression	87.22 ± 0.6206	87.01 ± 0.5672	86.74 ± 0.0049	86.68 ± 0.7788
Locally-deep SVN	86.52 ± 1.7445	86.44 ± 2.1287	86.23 ± 0.0043	86.75 ± 0.6856
Decision forest	87.96 ± 0.6110	89.09 ± 0.5473	85.57 ± 0.0031	88.91 ± 0.6783
Boosted decision tree	**89.79** ± **0.6625**	**90.34** ± **0.4984**	**90.97** ± **0.0024**	**91.61± 0.5655**
Boosted decision jungle	87.18 ± 1.0988	88.21 ± 1.1550	88.11 ± 0.0040	89.06 ± 0.8963

The AUC of the SAPS II score, the SOFA score and the chosen ML score are shown in figure 3. While it has been found that the inclusion of SAPS II score, compared with the SOFA score, tend to result in better overall ML performance when coupled with other variables, the opposite is seen when considering the ICU predictive scoring scores in isolation. Here, SOFA score did slightly better compared to SAPS II score by a difference of AUC 1.49%. Figure 3 shows that the ML model may improve the average prediction by 16.22%.

In addition to performance comparison, the relationship between risk groups for each severity score with the outcome of morbidity were also analysed. To do this, the conversion from SOFA score to mortality percentage was necessary. Here, the conversion followed the estimate of mortality risk based on studies by Vincent et al. [22] and Ferreira et al. [7]. Rate of mortality were 1.96%, 7.46%, and 90.58% patients were in the low, intermediate, and high risk ML score groups, respectively as shown in figure 4. Rate of mortality were 0.83%, 22.01%, and 43.95% patients were in the low, intermediate, and high risk based on the SAPS II score, respectively. While rate of mortality was 1.05%, 43.95%, and 55.0% patients were in the low, intermediate, and high risk based on the SOFA score, respectively. Here, it was demonstrated that the combined features present significant improvements to predictive accuracy and sensitivity compared to using SOFA score or SAPS II score alone and has shown good discriminating power for distinguishing patients who survived from those who died.

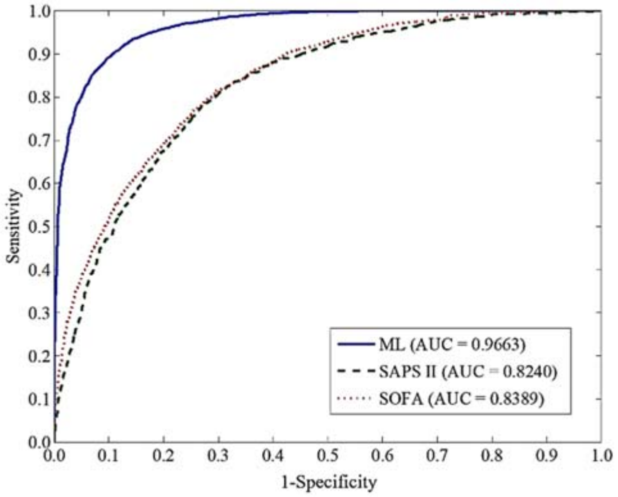

Figure 3. Recall for cross validation of n=10 folds without SMOTE (NS), SMOTE 100% (S100), SMOTE 300% (S300) and SMOTE 500% (S500) for (Blue).

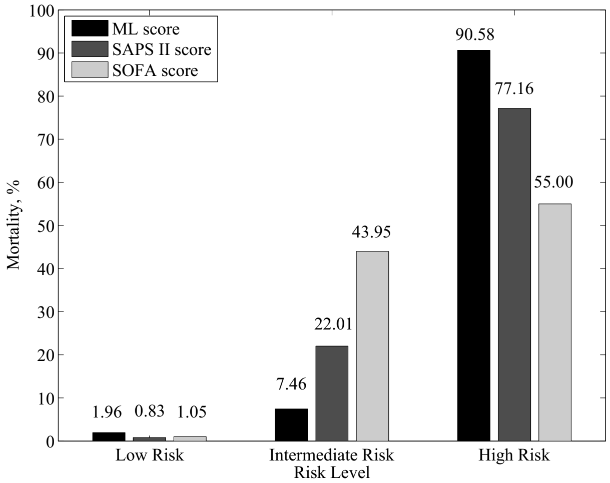

Figure 4. Risk groups for SOFA, SAPS II and ML score.

These results show that instead of developing locally-customised variants of the common scoring tools to improve calibration and variability in accuracy across various diseases and populations [9-13], ML allows us to leverage multidimensional analysis which includes parameters and data available in the Malaysian ICU database, such as physiological measurements, admission types etc., along with SOFA and SAPS II scores are able to improve accuracy of prediction mortality. This, in turn, will allow medical practitioners to make better medical decisions leading to more efficient resource management of limited resources which is critical in a public hospital ICU setting.

4. Future studies

This study includes patient records from multiple ICU's in Malaysia to reduce the potential of the algorithm overestimating since the training and the testing were executed on the partitions of the same data set. We are currently working on collecting more recent data which includes data from a wider range of hospitals. In the future, we would like to test and validate our algorithm with this data set to further increase the effectiveness of predictions and avoid performance variability in the algorithm.

5. Conclusion

The ML score proposed in this paper represents a noninvasive and objective risk-stratification tool that can be immediately determined at presentation to the ICU along with bedside diagnosis. It may be applied across populations as it 'learns' from data and not statistically modelled based on specific populations. This way, it is not required to localise the SAPS II and SOFA score in attempt to avoid variability. The ML score uses a combination of common clinical variables such as physiological measurements and admission types (see table 1) as a predictor of patient outcomes. A cross-validation was applied to the dataset of critically ill patients presenting to the ICU to overcome overfitting associated with traditional statistical methods. Various types of classification algorithms in machine learning was studied and it was found that the Decision Tree with SMOTE 500% provided the highest confidence in categorizing patients into two outcomes: death and survival. In addition, the ML score was found to predict mortality more accurately compared with the current ICU scoring system; SAPS II and SOFA scores. However, the incorporation of SAPS II and SOFA score in the dataset significantly improved the overall prediction performance of the ML score.

Acknowledgement

The authors would like to thank and acknowledge the contributions from all the doctors and nurses from the Intensive Care Units from Penang General Hospital, Kuala Lumpur Hospital, Sultanah Aminah Hospital and Tengku Ampuan Afzan Hospital. This study was financially supported by grants from University Malaysia Pahang (RDU160398) and Ministry of Higher Education Malaysia (FRGS/1/2017/TK03/UMP/03/1 and FRGS/1/2015/SKK02/UIAM/02/3).

References

[1] Vincent J L and Moreno R 2010 Clinical review: Scoring systems in the critically ill *Crit Care.* 14 207-15

[2] Rapsang A G and Shyam D C 2014 Scoring systems in the intensive care unit: A compendium *Indian J. Crit. Care Med.* 18 220-28

[3] Strand K and Flaatten H 2008 Severity scoring in the ICU: A review *Acta Anaesthesiol. Scand.* 52 467-78

[4] Granholm A, Møller M H, Krag M, Perner A and Hjortrup P B 2016 Predictive Performance of the Simplified Acute Physiology Score (SAPS) II and the Initial Sequential Organ Failure Assessment (SOFA) Score in Acutely Ill Intensive Care Patients: Post-Hoc Analyses of the SUP-ICU Inception Cohort Study *PLoS One* 11 e0168948

[5] Vincent J L, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, Reinhart C K, Suter P M and Thijs L G 1996 The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Med.* 22 707-10

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

[7] Ferreira F L, Bota D P, Bross A, Mélot C and Vincent J L 2001 Serial evaluation of the SOFA score to predict outcome in critically ill patients *JAMA* 286 1754-8

[8] Harrison D A, Brady A R, Parry G J, Carpenter J R and Rowan K 2006 Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. *Crit Care Med.* 34 1378-88

[9] Nassar A P J, Mocelin A O, Nunes A L, Giannini F P, Brauer L, Andrade F M and Dias C A 2012 Caution when using prognostic models: a prospective comparison of 3 recent prognostic models. *J Crit Care* 27 723.e1-e7

[10] Poole D, Rossi C, Latronico N, Rossi G, Finazzi S, Bertolini G and GiViTI 2012 Comparison between SAPS II} and SAPS 3 in predicting hospital mortality in a cohort of 103 Italian ICUs. Is new always better? *Intensive Care Med.* 38 1020-8

[11] Siontis G C, Tzoulaki I and Ioannidis J P 2011 Predicting death: an empirical evaluation of predictive tools for mortality *Arch Intern Med.* 171 1721-6

[12] Beck D H, Smith G B, Pappachan J V and Millar B 2003 External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: a multicentre study. *Intensive Care Med.* 29 249-56

[13] Apolone G, Bertolini G, D'Amico R, Iapichino G, Cattaneo A, De Salvo G and Melotti R M 1996 The performance of SAPS II in a cohort of patients admitted to 99 Italian ICUs: results from GiViTI (Gruppo Italiano per la Valutazione degli interventi in Terapia Intensiva) *Intensive Care Med.* 22 1368-78

[14] Metnitz P G H, Lang T, Vesely H, Valentin A and Le Gall J R 2000 Ratios of observed to expected mortality are affected by differences in case mix and quality of care *Intensive Care Med.* 26 1466-72

[15] Le Gall J R, Neumann A, Hemery F, Bleriot J P, Fulgencio J, Garrigues B, Gouzes C, Lepage E, Moine P and Villers D 2005 Mortality prediction using SAPS II: an update for French intensive care units *Crit Care.* 9 R645-52

[16] Metnitz B, Schaden E, Moreno R, Le Gall J, Bauer P and Metnitz P 2009 Austrian validation and customization of the SAPS 3 Admission Score *Intensive Care Med.* 35 616-22

[17] Moreno R P, Metnitz P G, Almeida E, Jordan B, Bauer P, Campos R A, Iapichino G, Edbrooke D, Capuzzo M and Le Gall J R 2005 SAPS 3 -- From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Med.* 31 1345-55

[18] Minne L, Eslami S, de Keizer N, de Jonge E, de Rooij S E and Abu-Hanna A 2012 Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment *Intensive Care Med.* 38 40-6

[19] Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P 2002 SMOTE: Synthetic Minority Over-sampling Technique *J. Artif. Intell. Res* 16 321-57

[20] Ong M E, Lee N C H, Goh K, Liu N, Koh Z X, Shahidah N, Zhang T T, Fook-Chong S and Lin Z 2012 Prediction of cardiac arrest in critically ill patients presenting to the emergency department using a machine learning score incorporating heart rate variability compared with the modified early warning score. *Crit Care.* 16 R108-20

[21] S Boughorbel , R Al-Ali and Elkum N 2016 Model Comparison for Breast Cancer Prognosis Based on Clinical Data *PLoS One* 11 1-15

[22] Vincent J L, de Mendonça A, Cantraine F, Moreno R, Takala J, Suter P M, Sprung C, Colardyn F and Blecher S 1998 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.* 26 1793-800