Health Outcome Predictive Modelling in Intensive Care Units

Chengqian Xian[1*], Camila P. E. de Souza[1†] and Felipe F. Rodrigues[1,2†]

[1*]Department of Statistical and Actuarial Sciences, University of Western Ontario, 1151 Richmond Street, London, N6A 5B7, Ontario, Canada.

[2]School of Management, Economics, and Mathematics, King's University College at Western University, 266 Epworth Avenue, London, N6A 2M3, Ontario, Canada.

*Corresponding author(s). E-mail(s): cxian3@uwo.ca;
Contributing authors: camila.souza@uwo.ca; frodrig7@uwo.ca;
†These authors contributed equally to this work.

Abstract

The literature in Intensive Care Units (ICUs) data analysis focuses on predictions of length-of-stay (LOS) and mortality based on patient acuity scores such as Acute Physiology and Chronic Health Evaluation (APACHE), Sequential Organ Failure Assessment (SOFA), to name a few. Unlike ICUs in other areas around the world, ICUs in Ontario, Canada, collect two primary intensive care scoring scales, a therapeutic acuity score called the "Multiple Organs Dysfunctional Score" (MODS) and a nursing workload score called the "Nine Equivalents Nursing Manpower Use Score" (NEMS). The dataset analyzed in this study contains patients' NEMS and MODS scores measured upon patient admission into the ICU and other characteristics commonly found in the literature. Data were collected between January 1st, 2015 and May 31st, 2021, at two teaching hospital ICUs in Ontario, Canada. In this work, we developed logistic regression, random forests (RF) and neural networks (NN) models for mortality (discharged or deceased) and LOS (short or long stay) predictions. Considering the effect of mortality outcome on LOS, we also combined mortality and LOS to create a new categorical health outcome called LMClass (short stay & discharged,

short stay & deceased, or long stay without specifying mortality outcomes), and then applied multinomial regression, RF and NN for its prediction. Among the models evaluated, logistic regression for mortality prediction results in the highest area under the curve (AUC) of 0.795 and also for LMClass prediction the highest accuracy of 0.630. In contrast, in LOS prediction, RF outperforms the other methods with the highest AUC of 0.689. This study also demonstrates that MODS and NEMS, as well as their components measured upon patient arrival, significantly contribute to health outcome prediction in ICUs.

Keywords: Intensive Care Units, Health outcome, Predictive modelling, MODS, NEMS

1 Introduction

The Intensive Care Unit (ICU) is a unique hospital department, providing the highest level of treatment for a hospital's highest acuity patients. It is an intrinsically technological environment where each patient generates thousands of data points per day, and data-driven management applied to ICU allows not only an evaluation of ICU performance but has other implications, including better planning of scarce resources and transition of care and discharge [1, 2].

ICU scoring system is an essential tool to describe the severity of patients' illnesses to improve clinical decision-making and predict patients' health outcomes [3–7]. Existing scoring systems to assess patient illness severity on admission to ICU include the Acute Physiology and Chronic Health Evaluation II and its variations (APACHE II) [8] as well as the Simplified Acute Physiology Score (SAPS II) [9]. In addition to describing the severity of a patient's disease, Multiple Organ Dysfunction Score (MODS) [10] and Sequential Organ Failure Assessment Score (SOFA) [11] were also developed to specifically evaluate patients' organ function or determine the rate of organ failure. The MODS score (see the full list of its components in Appendix Table A6) is scaled from 0 to 24, and it is constructed from six organ systems and demonstrates a strong correlation with the risk of ICU mortality [10]. The Nine Equivalents Nursing Manpower Use Score (NEMS) [12] (see the full list of its components in Appendix Table A7) was developed from the Therapeutic Intervention Scoring System (TISS) [13] to measure the nursing workload in ICU. NEMS is based on nine life support interventions ranging from 0 to 56, and has been validated in an adult 30-bed medical-surgical ICU in a tertiary care university hospital. Its good agreement is further confirmed with TISS-28 [14].

In Ontario, Canada, hospitals collect MODS and NEMS for reporting purposes, but they lack the necessary measurements to calculate widely used severity scores like APACHE and SAPS [15]. Limited recent studies have explored the relationship between MODS, NEMS, and the health outcomes of ICU patients in Ontario. Notably, two main studies conducted by Kao et al. [16] and Rodrigues [15] provide valuable insights. In their study, Kao et al.

developed a logistic regression model to predict patient mortality using MODS, NEMS, and general patient characteristics, such as age, based on a data set comprising 8, 822 patients collected from Ontario between January 1, 2009, and November 30, 2012. Rodrigues expanded on Kao et al.'s work in two ways using more recent data collected between January 1, 2015, and December 31, 2016, which included 4, 758 patients. First, Rodrigues enhanced the mortality prediction models by incorporating the components of MODS and NEMS, not only through logistic regression but also utilizing supervised machine learning methods like random forests and neural networks. Additionally, Rodrigues focused on developing models for predicting length of stay (LOS) in the ICU, as it significantly influences ICU resource planning due to the strong correlation between ICU costs and LOS [17]. The key distinction between these two studies lies in their research focus. Rodrigues primarily compares the performance of multiple statistical and supervised machine learning algorithms in predicting mortality and LOS, while Kao et al. primarily investigate statistically significant predictors in distinguishing patients at low and high risk of mortality.

In this work, we expand the work of [16] by adding the components of MODS and NEMS as predictors. We then consider logistic regression and compare it with two other common machine learning methods (random forests and neural networks) for mortality or LOS predictions. We validate the work of [15] using a larger data set with 15, 350 patients collected from the same ICUs between Jan. 1, 2015, and May 31, 2021. Our main research objectives are: (i) investigate the significance of MODS and NEMS along with their components in both mortality and LOS predictions; (ii) analyze the quantitative effect of patient general characteristics and the admission characteristics (e.g., MODS and NEMS) via regression models; (iii) construct a new categorical outcome, LMClass (LOS-Mortality Class), to combine mortality and LOS, and fit multinomial regression models, random forests, neural networks for its prediction.

As we will demonstrate and discuss, our proposed models with MODS and NEMS components added as predictors significantly improve the performance of mortality and LOS predictions. Compared with [15], our second objective fills the gap in the interpretation of risk factors on health outcomes predictive modelling and provides a better understanding of a predictive model. More sophisticated methods like machine learning algorithms may perform better in classifying health outcomes. However, their complex model structures make it harder to understand and may lose interpretation power [18]. Additionally, the motivation for analyzing LMClass comes from the "endogeneity" of mortality in ICU length of stay prediction [15, 19]. For instance, deceased patients may have shorter or longer LOS than their discharged counterparts. Ignoring this endogenous effect may cause bias in LOS prediction [20]. Combining these two outcomes may provide a more comprehensive and useful assessment of patient outcomes [21]. We, therefore, suggest a new categorical health outcome, LMClass, with three categories: short stay & discharged; short stay &

deceased; and long stay without specifying mortality outcomes. We will discuss the definition of a prolonged LOS in Section 3.1. In this composite outcome, we focus on mortality outcomes in patients with short LOS for short-term allocation and planning of ICU resources.

The rest of the paper is structured as follows. We provide a general literature review on health outcome predictive modelling in ICU in Section 2. Section 3 presents the material and the statistical methodology. Then, Section 4 describes our analysis results. Finally, the conclusion and discussion are presented in Section 5.

2 Literature review

With advances in information technology and data science, statistical models and machine learning methods have been applied to ICU data for mortality and LOS predictions [1, 22–25]. In what follows, we present a comprehensive literature review of mortality prediction in Section 2.1 and LOS prediction in Section 2.2.

2.1 Mortality prediction

ICU mortality prediction, involving the classification of patients as either discharged or deceased, is commonly approached as a binary classification problem. Extensive research has been conducted for ICU mortality prediction. For comprehensive insights in this field, systematic reviews by Fusaro et al. [26] and Keuning et al. [27] are valuable resources.

Logistic regression combined with the likelihood ratio test (LRT) is widely used to predict patient mortality in ICU [16, 28–36]. In a recent study [37], logistic regression was employed to detect risk factors associated with ICU survival during the COVID-19 pandemic. As one of the most commonly used generalized linear models, logistic regression has an excellent interpretation power via odds ratio [38, 39], which helps quantitatively describe the impact of each predictor in mortality risk [18].

Machine learning algorithms are widely acknowledged as alternatives to logistic regression in various domains, including ICU mortality prediction. These algorithms encompass a range of methods, such as decision trees, support vector machines, k-nearest neighbors, random forests, super learners, boosting, and neural networks, among others [15, 40–44]. Furthermore, there exists an extensive body of literature that explores the application of machine learning methods to predict ICU mortality outcome specifically for COVID-19 patients [45–49]. When reviewing these studies, it is essential to consider several unique characteristics of the analyzed ICU data. These factors include the country or region from which the data originates, the patient population under study (e.g., adults or children, patients with severe pneumonia or heart disease), and whether the data is sourced from a single center or multiple centers. By taking these factors into account, researchers can better contextualize and interpret the findings from these diverse studies.

2.2 LOS prediction

When reviewing the literature on ICU length-of-stay (LOS) prediction, we can categorize the existing studies into three main groups: typical regression analysis, binary classification, and survival regression. In typical regression analysis, two common approaches are employed: multiple linear regression (MLR) and regression using machine learning techniques. For binary classification, the first step is to define what constitutes a prolonged LOS. Researchers need to establish a threshold or criteria for defining a long stay in the ICU. Once this definition is established, classification methods are applied to predict whether a patient will have a prolonged LOS based on the available features and predictors. Survival regression analysis is another approach used in LOS prediction, specifically designed to handle censoring in the data collection process. It is worth noting that different scoring systems may be employed in various studies, depending on the geographical location of the ICUs and the specific context of the research. Recent comprehensive reviews on LOS prediction can be also found in [50–52].

Zimmerman [7] developed an MLR procedure using APACHE IV to estimate ICU stay using data across ICUs in the United States. They showed that the accuracy and utility of the predictions based on the APACHE IV model were unsatisfactory. In [53], researchers improved patient LOS prediction by firstly optimizing a threshold for a prolonged stay and building a multivariate linear regression with the severity score information on day five, achieving a better prognosis than that based on ICU day one information alone. Similar studies where the MLR was applied can be found in [28, 54, 55]. Regression models based on machine learning are also widely considered and built, which include support vector regression [56], gradient boosting regression [57], random forests regression [58, 59].

Sometimes, clinical practitioners are also interested in the binary classification for LOS prediction (long-stay or short-stay), and therefore classification methods including logistic regression, support vector machine, random forests, and neural networks were also implemented in predicting prolonged LOS or short LOS [15, 56, 60, 61]. Neural networks were developed as predictive instruments for ICU LOS for the first time in [60]. Defining prolonged LOS as a stay greater than two days, they found that the neural networks model performed well with an area under the receiving operating characteristic curve (AUC) of 0.6960 in the validation set. In [56] and [15], the authors performed similar work on LOS prediction by applying machine learning methods but based on different severity scores (SOFA score in [56] while MODS in [15]).

Recently, survival regression analysis is also conducted for LOS prediction, where the time of ICU stay is considered a survival time response to correct for censoring. The AFT model with Weibull distribution was developed in [15] for short-term capacity planning in ICUs from Ontario, Canada. The Weibull AFT model was also applied in [62, 63] to investigate the effect of predictors on LOS for COVID-19 patients in the UK. Authors in [62] further built the log-normal AFT compared to the one with Weibull distributional assumption.

The Cox PH model was developed in [64] to analyze the effect of severity scores, SOFA, on LOS for COVID-19 patients in India.

3 Material and methods

3.1 Data source and data management

Our research is a retrospective study conducted at two teaching hospital ICUs in southwestern Ontario, which specialize in the care of various patient populations, including neurosurgical, cardiovascular surgery, and transplantation patients. Data were collected from Jan. 1, 2015, to May 31, 2021 and stored in four separate data sets called MODS, NEMS, Source and Awaiting Transfer. MODS is the data set containing the MODS score along with its components measured upon patient admission to ICU. NEMS is another important set containing patients' NEMS scores and their components measured daily in ICU. ICU discharge time and destination are also provided in MODS and NEMS data sets. The Source set includes de-identified patient general characteristics (e.g., age, sex) and admission characteristics (e.g., admission source, admission diagnosis, patient category, referring service). The last data set, Awaiting Transfer, provides the admission time and the awaiting transfer discharge start date time, both of which were used to calculate the clinical LOS.

Since we have four separate data sets, several new variables were created in each set before merging them into a single data set. In the MODS set, we created *Mortality* as a binary response which can be constructed from discharge destination: 1 if the patient is deceased at the end of the ICU stay, otherwise 0 for being discharged alive. Besides, we calculated the *total LOS*, defined as the period between patients' admission to and exit from ICU, which is used to detect extreme values of stay in data cleaning procedure as done in [15]. In the Source set, we edited the admission source and admission diagnosis by combining some of their categories in the same way proposed in [15]. For admission sources, we kept the Emergency Department, Operating Room, and Unit/Ward/Stepdown while combining the other levels to Outside Hospital/Other. For admission diagnosis, we kept the Cardiovascular/Cardiac/Vascular, Gastrointestinal, Neurological, Respiratory, and Trauma while combining other levels to Other. In the Awaiting Transfer set, we calculated the *clinical LOS*, defined as the period between patient admission to ICU and the physician's disposition decision (i.e., transfer or discharge). Then the prolonged LOS called *IsLong* is defined as a stay longer than five days, based on the empirical distribution of LOS as discussed in Section 4.1. In other words, *IsLong* takes the value of 1 if clinical LOS > 5 days and 0 otherwise. Besides, *LMClass*, a categorical response with three levels, was also created by combining *Mortality* and *IsLong*: short stay & discharged, short stay & deceased, or long stay without specifying mortality outcomes.

To build predictive models for each health outcome, we need to combine these four separate data sets into a single one. Patients' ID and admission time can link these four data sets. We first extracted the MODS score with

its components from the MODS set and used patients' IDs and admission time to merge the NEMS score with its components on admission day in the NEMS set. To obtain the clinical LOS, patient characteristics, and admission characteristics, we merged the latest combined data set with the Awaiting Transfer set and the Source set, resulting in a single data set with 15,474 cases. Similar to [15], some cases with large total LOS (≥ 60 *days*, 90 cases), unknown or missing sex (15 cases), and unusual age (≥ 110 *years*, 19 cases) were removed from our merged data set for further analysis, resulting in a finalized data set with 15,350 cases. A flow chart of this process is provided in Figure 1.

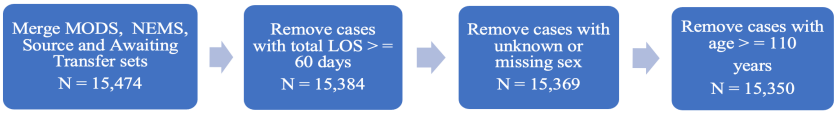

Fig. 1: Flow chart of data cleaning

3.2 Statistical analysis

The finalized data set ($N = 15,350$) was split into a training set ($N = 10,745$) and a validation set ($N = 4,605$) with a ratio of 7:3. Each proposed model was built on the training set and validated on the validation set. Figure 2 presents a flow chart of ICU health outcome predictive modelling. To present more details of our methodology and assure reproduciblility, we fill in the scorecards suggested by [65] for mortality or LOS prediction (Table A1 in Appendix A) and LMClass prediction (Table A2).

Logistic and multinomial regression are two widely used generalized linear models for modelling binary and multi-class responses, respectively [18]. As we illustrated in our objectives, we aim to apply the odds ratio with respect to a unit change from the risk factor to describe the corresponding effect on the health outcome. For example, in mortality prediction, considering a fitted logistic regression model on a series of risk factors including MODS with a positive estimated regression coefficient denoted by $\hat{\beta}$, we can describe as follows: with other risk factors unchanged, an ICU patient with one unit increase in MODS score will increase the odds of death by $100\big(\exp\big(\hat{\beta}\big) - 1\big)\%$ [18].

Random forests (RF), proposed in [66], is another popular model for classification by constructing a multitude of decision trees. The two most critical parameters in the random forests model are the number of trees to be built and the number of variables randomly sampled as candidates at each split. In the R package *RandomForest* [67], these parameters are represented by function arguments *ntree* and *mtry*, respectively. In our data set, there are 23 predictors, so *mtry* can be $1, 2, ...,$ or 23. For the number of trees, we chose from

100, 500 and 1000. Combining both *ntree* and *mtry*, we have 69 (i.e., 23×3) alternative models built on the training set using all predictors available. To analyze the black-box mechanisms of random forests, one of the most efficient variable importance measures, mean decrease accuracy (MDA) introduced in [68], can be applied, which is a method of computing the predictor importance on permuted out-of-bag samples based on the mean decrease in the accuracy. In other words, if MDA is high for a predictor, this predictor is important. Visualization of MDA for all predictors is provided after fitting an RF model in the same R package, *RandomForest*.

Neural networks (NN) were built based on the resilient back-propagation with weight backtracking algorithm proposed by Riedmiller M. in 1994 [69]. Before modelling, we conducted data preprocessing for numeric predictors (e.g., age) and ordinal categorical predictors (e.g., components of MODS) by min-max normalization and for nominal categorical predictors (e.g., admission diagnosis) by a one-hot encoding scheme [70]. The two most important parameters of the NN are the number of hidden layers and the number of neurons on each layer. We consider one or two hidden layers with one to five neurons, and as a result, we need to find the optimal NN model from 30 (i.e., $5 + (5 \times 5)$) alternative combinations of different numbers of hidden layers and neurons. All statistical analyses were performed using software R version 4.2.1.

In what follows, the evaluation metrics for the classification are demonstrated. In mortality and LOS binary predictions, we assessed model discrimination performance by AUC from the receiver operating characteristic (ROC) curve, sensitivity (Sen, also called "recall"), specificity (Spe), accuracy (Acc), Matthews correlation coefficient (MCC), positive predictive value (PPV, also called "precision"), negative predictive value (NPV) and F1 score. In LMClass prediction, which is a three-class classification problem, we calculate the accuracy (i.e., the overall percentage of cases correctly classified), the balanced accuracy (i.e., the average of recalls from each class) and the Kappa statistic [71] to evaluate the performance of each proposed model. In binary classification, we use the AUC as a criterion for parameter optimization in RF and NN models, while we use the Kappa statistic for LMClass classification. As discussed in [72], AUC evaluates the overall diagnostic performance of a binary classification and helps select the optimal threshold for determining the presence or absence of a specific health outcome. Furthermore, the Kappa statistic in multi-class classification is an appropriate metric to account for class imbalance [73].

4 Results

In this section, we first present the descriptive analysis results of our data set, and then the results regarding mortality, LOS, and LMClass prediction, respectively. We also quantitatively elaborate on how MODS and NEMS affect the prediction of health outcomes via the odds ratio (i.e., the relative risk ratio) from regression-based models.

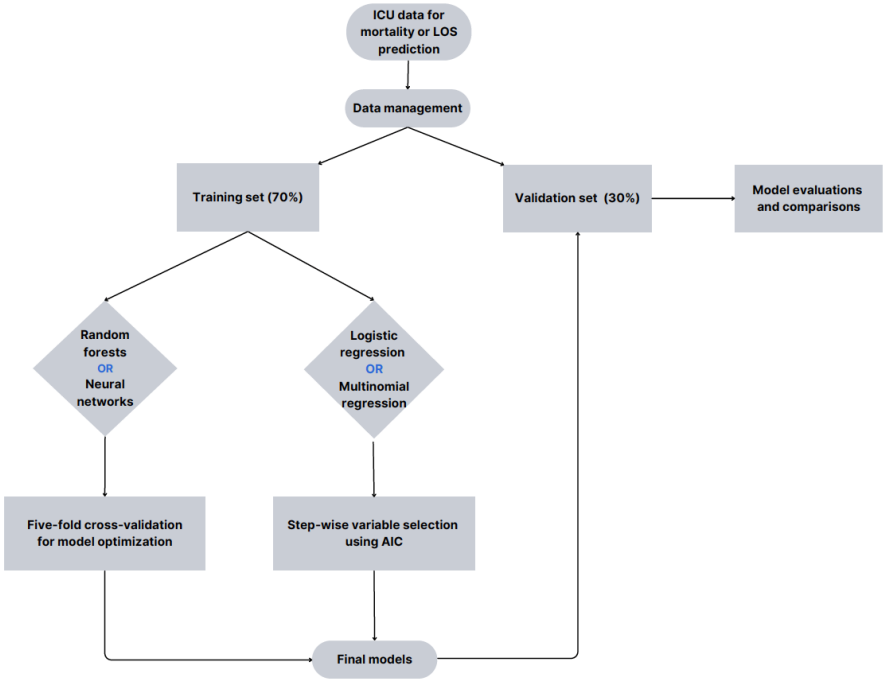

Fig. 2: Flow chart of ICU health outcomes predictive modelling

4.1 Descriptive analysis

In our dataset, 11,963 admitted patients were discharged alive when exiting the ICU while 3,387 patients died. Most patients stayed less than 5 days in ICU, accounting for 70% of the study population. Table 1 shows the general characteristics of patients (excluding the components of MODS and NEMS) in the training and validation sets. The median with interquartile interval (IQI) was presented for numeric variables, and for categorical variables, raw counts and percentages were presented. We can observe that the median values of MODS and NEMS scores are the same in the training and validation sets (5 points with IQI 3-7 and 32 points with IQI 27-39, respectively). The mortality rate in training is 21.83%, 0.78% lower than that in the validation set. Median clinical LOS in training and validation is 2.545 days (IQI 1.082-5.776) and 2.537 days (IQI 1.115-5.878), respectively.

Figure 3 shows the histogram with the estimated density function (the red curve) of clinical LOS. There are only a few cases with clinical LOS longer than 20 days and most of cases have clinical LOS between 0 and 4 days, resulting in a right-skewed distribution. Specifically, 30% of ICU patients in the data set stayed longer than 5 days and 15% longer than 10 days. Teaching hospitals in Ontario define a prolonged ICU LOS as longer than 21 days [15], which accounts for 4.1% in our data set. Therefore, we need to adjust the definition of

a prolonged LOS to preserve clinical significance while avoiding an extremely imbalanced data set. We consider a stay longer than 5 days (i.e., the 70th percentile) as a prolonged stay in our data set. First, patients may require 2-day stay in ICU for routine postoperative monitoring [74, 75]. Second, most standard classification machine learning methods including random forests face great challenge in presence of imbalanced data [76]. It is important to note that different studies may define prolonged LOS using different thresholds, for example, 3-day [77], 5-day [78, 79], 7-day [80], 8-day [81], and 21-day [82].

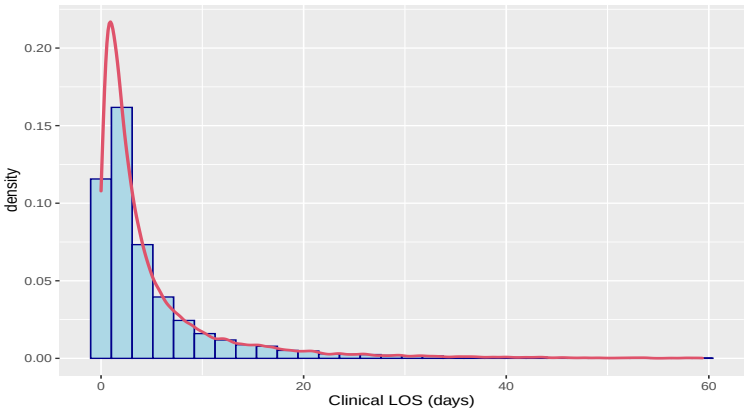

Fig. 3: Histogram plot with estimated density (red curve) of clinical LOS.

4.2 Mortality prediction

The best RF model in mortality prediction is with $mtry = 3$ and $ntree = 1000$. In other words, the number of variables randomly sampled as candidates at each split on the tree is three, and 1000 trees were built to construct the forests. The best NN model has one hidden layer on which there are two neurons.

Performance of logistic regression, RF and NN on the validation set for mortality prediction is shown in Table 2. ROC curves for all models in the validation set are also presented in Figure 4 which shows no big difference among different models. Logistic regression outperforms RF and NN with the highest scores in AUC (0.795), accuracy (0.705), F1 score (0.532) and PPV (0.415). RF performs the best in achieving the highest sensitivity of 0.748 and NPV of 0.904. Logistic regression has the second highest sensitivity (0.743), while NN provides a relatively lower sensitivity (0.732). The MCC values are low, ranging from 0.364 to 0.373 and the F1 scores are better, ranging from 0.527 to 0.532.

In mortality prediction, the F1 scores from the proposed models are slightly larger than 0.5, which, one may think, is close to a random guess. The F1 score in binary classification takes into account both PPV (i.e., precision)

Table 1: Characteristics of ICU admitted patients in the training and validation sets.

Variables	Training	Validation
MODS Score (median [interquartile interval] [1])	5 [3, 7]	5 [3, 7]
NEMS Score (median [interquartile interval])	32 [27, 39]	32 [27, 39]
Mortality, yes, (N, [%] [2])	2346 [21.83%]	1041 [22.61%]
Total LOS, days, (median, [interquartile interval])	3.361 [1.671, 6.913]	3.372 [1.645, 6.885]
Clinical LOS, days, (median, [interquartile interval])	2.545 [1.082, 5.776]	2.537 [1.115, 5.878]
Age, years, (median, [interquartile interval])	63.39 [51.14, 73.46]	63.55 [50.94, 73.78]
Sex, Male, (N, [%])	6269 [58.34%]	2690 [58.41%]
Patient Category		
Medical, (N, [%])	6769 [63.00%]	2855 [62.00%]
Surgical, (N, [%])	3976 [37.00%]	1750 [38.00%]
Admission Source		
Emergency Department, (N, [%])	3740 [34.81%]	1551 [33.68%]
Operating Room, (N, [%])	2045 [19.03%]	922 [20.02%]
Unit/Ward/Stepdown, (N, [%])	2214 [20.60%]	960 [20.85%]
Outside Hospital/Other, (N, [%])	2746 [25.56%]	1172 [25.45%]
Admission Diagnosis		
Cardiovascular/Cardiac/Vascular, (N, [%])	1754 [16.32%]	750 [16.29%]
Gastrointestinal, (N, [%])	988 [9.19%]	417 [9.06%]
Neurological, (N, [%])	1473 [13.71%]	561 [12.18%]
Respiratory, (N, [%])	3111 [28.95%]	1336 [29.01%]
Trauma, (N, [%])	832 [7.74%]	370 [8.03%]
Other, (N, [%])	2587 [24.08%]	1171 [25.43%]

[1]for numeric variables, the median with interquartile interval was presented;

[2]for categorical variables, we presented raw count (N) and the percentage (%) of each category indicated by an indentation.

and sensitivity (i.e., recall) to provide a more robust evaluation in case of data imbalance. In our predictive modelling for patient mortality, we obtained a lower PPV, around 0.4, while a higher sensitivity, around 0.7, from each proposed model. Therefore, the F1 score is around 0.5. In principle, an F1 score of 0.5 indicates a lower prediction performance. However, such an F1 score resulting from a higher sensitivity still carries meaning for this type of application. For example, since we define death as the outcome of interest, a higher sensitivity means the model correctly detects most patients who will actually die. Moreover, if our model predicts a patient's death, the patient is still likely to survive based on our PPV values. In contrast, due to our high NPV values, a survival prediction is likely a true negative, i.e., the patient survives.

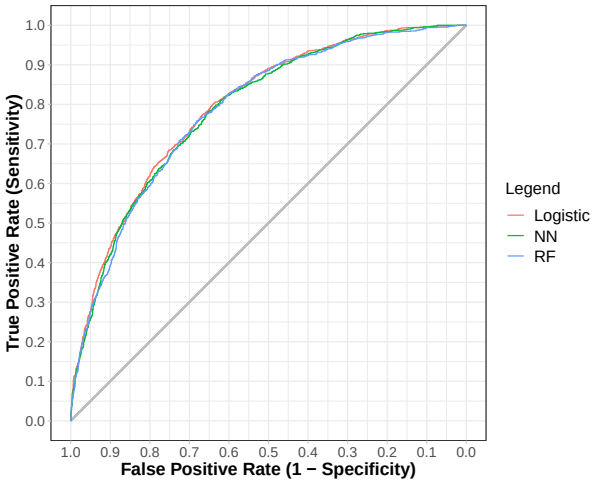

Fig. 4: ROC curves for mortality prediction in the validation set.

Red: logistic regression; Green: random forests; Blue: neural networks

Table A3 in the Appendix A shows the results of selected predictors for logistic regression based on AIC. Those selected predictors include six components of MODS (Haematologic, Hepatic, Renal, Cardiovascular, Neurologic and Respiratory), NEMS score, six components of NEMS (Basic Monitoring, Intracranial Pressure Monitor, Dialysis, Intra-Aortic Balloon Pump, Other Interventions Within this Unit and Interventions Outside this Unit), Age, Sex, Patient Category, Admission Source and Admission Diagnosis. The odds ratio and its 95% confidence interval (CI) were also calculated for each selected predictor. We find that MODS score is not selected in the optimal model but its six components are. However, NEMS is selected in the model and has an odds ratio of 1.07 (95% CI = [1.06, 1.08]), which indicates that one point increase in NEMS score would increase the relative risk of mortality (i.e., death) by 7% (95% CI = [6%, 8%]), holding the other covariates fixed.

Table 2: Performance of logistic regression, RF and NN on the validation set for mortality prediction.

Model	AUC	95% CI	Cutoff	Sen	Spe	Acc	MCC	F1	PPV	NPV
Logistic	0.795	(0.780, 0.810)	0.206	0.743	0.694	0.705	0.372	0.532	0.415	0.902
RF	0.788	(0.773, 0.803)	0.186	0.748	0.690	0.703	0.373	0.532	0.413	0.904
NN	0.789	(0.774, 0.804)	0.221	0.732	0.695	0.703	0.364	0.527	0.412	0.899

AUC: area under the curve; 95% CI: 95% confidence interval for AUC; Cutoff: optimal threshold for determining mortality obtained via Youden's J statistic [83]; Sen: sensitivity; Spe: specificity; Acc: accuracy; MCC: Matthews correlation coefficient; F1: F1 score; PPV: positive predictive value; NPV: negative predictive value.

The significant variables returned by the regression model and the important variables from random forests have a high degree of agreement. In logistic regression, the likelihood ratio test (LRT) was applied to all the selected predictors to assess their significance [18], and the corresponding p-values were reported in the A3. Except for one of the NEMS components, the Intra-Aortic Balloon Pump, we find that all other selected predictors are statistically significant for mortality prediction. We present a visualization of predictor importance using MDA from the RF model in the left plot of 5. The plot shows that NEMS, MODS, ICU admission source, age, neurological level, patient category and ICU admission diagnosis are the seven most important predictors with an MDA higher than 30%. In addition, these seven predictors are all significant predictors in the logistic regression model for mortality prediction.

Practitioners may be interested in the sensitivity (i.e., recall) of the fitted model for mortality prediction. Specifically, they are concerned about the proportion of correct prediction for those patients who deceased at the end of ICU stay. In mortality prediction, the sensitivity coming from RF is 0.748, meaning that it works relatively well in predicting the mortality outcome in those patients who deceased in the end and 74.8% could be correctly predicted. NPV is also an important index in mortality prediction. RF has the highest NPV of 0.904, meaning that if we predict that someone will be discharged alive at the end of the stay, they would likely be discharged with a probability of 90.4%.

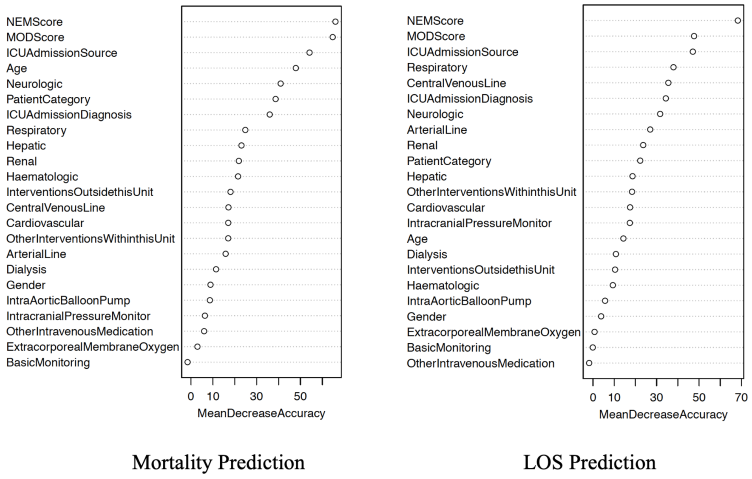

Fig. 5: Importance of predictors based on RF model in mortality (left) and LOS (right) predictions.

4.3 LOS prediction

In LOS prediction, the optimal RF model has the same parameter structure as that in mortality prediction where three predictors were randomly sampled at each split on the tree and 1000 trees were built. The best NN model has one hidden layer with one neuron. This simplest NN model was also presented and investigated as the best model in [84] for survival prediction in the ICUs, which is consistent with the empirical analysis result that reducing the complexity of a neural network structure may provide a better performance of prediction for health outcomes in ICUs [85, 86].

Table 3 presents the evaluation measures of all models for LOS prediction in the validation set. The AUC values of all models are lower than those in mortality prediction by around 10%. The MCC values ranging from 0.247 to 0.251 are very low while the F1 scores ranging from 0.503 to 0.508 are still acceptable. On the whole, logistic regression and RF outperform NN. Specifically, logistic regression has the highest score in sensitivity (0.673), MCC (0.251) and NPV (0.820), while RF has the highest score in AUC (0.689), specificity (0.622), accuracy (0.630) and PPV (0.411). The corresponding ROC curves were shown in Figure B1 in Appendix B and once again, no big difference among models can be seen from the plot.

In LOS prediction, sensitivity is one of the most important indices for the choice of models. Logistic regression has the highest sensitivity of 0.673, indicating that 67.3% of patients who stayed more than 5 days could be correctly predicted. The negative predictive values (NPV) from logistic regression is also the highest (0.82), meaning that if we predict one stays less than 5 days, they would probably stay less than 5 days with a probability of 0.82.

Table A4 in Appendix A contains information of the selected predictors for logistic regression in LOS prediction. MODS score with its two components (Cardiovascular and Respiratory), NEMS score with its five components (Central Venous Line, Arterial Line, Intracranial Pressure Monitor, Dialysis and Interventions Outside this Unit), Sex, Admission Source and Admission Diagnosis are selected in the best model. The odds ratios for MODS and NEMS are 1.04 (95% CI = [1.02, 1.06]) and 1.04 (95% CI = [1.04, 1.05]), respectively. This indicates that one point increase in MODS score or in NEMS score would increase the relative risk of staying more than 5 days by 4%, holding the other predictors fixed. Furthermore, from the right plot of Figure 5, we can see NEMS and MODS are both important predictors in prolonged LOS prediction, which is consistent with results of LRT for testing the significance of MODS and NEMS based on logistic regression.

4.4 LMClass prediction

In LMClass prediction, only NN models with one hidden layer reliably converged. Therefore, we considered one to ten neurons within the hidden layer and found that three neurons yielded the highest Kappa statistic. For the RF

Table 3: Performance of logistic regression, RF and NN on the validation set for LOS Prediction.

Model	AUC	95% CI	Cutoff	Sen	Spe	Acc	MCC	F1	PPV	NPV
Logistic	0.681	(0.665, 0.698)	0.275	0.673	0.604	0.624	0.251	0.508	0.408	0.820
RF	0.689	(0.673, 0.705)	0.272	0.650	0.622	0.630	0.247	0.503	0.411	0.814
NN	0.682	(0.665, 0.698)	0.299	0.672	0.604	0.624	0.250	0.508	0.408	0.819

AUC: area under the curve; 95% CI: 95% confidence interval for AUC; Cutoff: optimal threshold for determining mortality obtained via Youden's J statistic [83]; Sen: sensitivity; Spe: specificity; Acc: accuracy; MCC: Matthews correlation coefficient; F1: F1 score; PPV: positive predictive value; NPV: negative predictive value.

model fitting, the optimal number of random predictors at each split and the optimal number of trees built are 9 and 500, respectively.

Table 4 presents each overall accuracy, balanced accuracy and the Kappa statistic from the fitted multinomial regression, RF and NN models on the validation set. Multinomial regression, with a higher accuracy of 0.630, slightly outperforms the NN model, which has an accuracy of 0.628. However, the NN model results in the highest Kappa statistic of 0.303 and the highest balanced accuracy of 0.5. All three models have a balanced accuracy of around 0.5 and a Kappa statistic of around 0.3, indicating room for improvement in predicting LMClass with imbalanced categories.

The information of selected predictors in the multinomial regression model for LMClass prediction is provided in Table A5 in Appendix A . In multinomial regression, we set the baseline class to be short stay & discharged, and the odds ratio with 95% CI for each selected predictor was collected for another two classes (e.g. short stay & deceased and long stay without specifying mortality outcomes) with respect to the baseline. The odds ratios of MODS for the class short stay & deceased and the class long stay without specifying mortality outcomes are 1.40 (95% CI = [1.34, 1.46]) and 1.15 (95% CI = [1.11, 1.19]), respectively. This means if one point increase in MODS score would increase the relative risk of short stay & discharged over short stay & deceased and long stay without specifying mortality outcomes by 40% (95% CI = [34%, 46%]) and 15% (95% CI = [11%, 19%]), respectively. Similarly, the odds ratios of NEMS for short stay & deceased and long stay without specifying mortality outcomes are 1.11 (95% CI = [1.10, 1.13]) and 1.07 (95% CI = [1.06, 1.08]), respectively. This means that one point increase in NEMS score would increase the relative risk of short stay & discharged over short stay & deceased and long stay without specifying mortality outcomes by 11% (95% CI = [10%, 13%]) and 7% (95% CI = [6%, 8%]), respectively, holding the other predictors fixed. As a reference, importance of predictors based on MDA is visualized in Figure B2 in the Appendix B which shows that both NEMS and MODS are important in LMClass prediction.

Table 4: Performance of multinomial regression, RF and NN models on the validation set for LMClass prediction.

Model	Accuracy	Balanced accuracy	Kappa
Multinomial	0.630	0.499	0.299
RF	0.619	0.494	0.290
NN	0.628	0.500	0.303

5 Conclusions and Discussion

In this work, we developed several models for health outcomes prediction in intensive care units. Compared with [16], adding the components of MODS

and NEMS in the logistic regression for mortality prediction has an improvement of 3.5% in AUC values in the validation set (see Table 5). This study also demonstrates that MODS and NEMS with their components measured upon patient arrival significantly contribute to health outcome prediction in ICUs. In mortality prediction, achieving the highest sensitivity and NPV, RF outperforms logistic regression and NN, but logistic regression achieves the highest AUC. In LOS prediction, no big difference in the performance appears in the logistic regression and NN model. In practice, we need to evaluate the pros and cons of each model, and choose the best according to the type and goals of the analysis. For example, if we are concerned about correctly predicting the mortality outcomes among all the ICU patients, logistic regression is our first choice, while we would choose RF if we emphasize on prediction accuracy among the deceased patients. Explanation power may also play a role, especially with respect to the implications of such predictions. As an example, the predictors of long stays may help inform capacity planning and resource scheduling.

Furthermore, in terms of the definition of prolonged stay in the ICU, random forests and neural networks have greatly improved LOS prediction when we cut the short and long LOS at 5 days instead of 7 or 21 days as in [15]. A comparison on AUC values in LOS prediction between previous works and our study is provided in Table 6. However, as in [15] for LOS prediction, we find it is harder to classify a short or long stay than to detect mortality status. The underlying reason could be the definition of prolonged LOS as a binary health outcome. To improve the prediction accuracy, survival models can be developed for LOS prediction, and in this scenario LOS can be considered as a continuous time-to-event response.

A trade-off between interpretation power and accuracy of prediction usually exists in predictive modelling. Logistic and multinomial regression models provide an interpretation for quantitative relationships between predictors and health outcomes using odds (i.e., relative risk). Compared with regression-based models, RF provides qualitative relationships using MDA, while NN is a black box whose statistical theoretical justifications are still under investigation in different frameworks [87, 88].

To our best of knowledge, we are the first to combine mortality with prolonged LOS to construct a new categorical health outcome and develop MODS and NEMS based predictive models for its prediction. In our expectation, more complexity occurs in this three-level outcome, making it more challenging to achieve high prediction accuracy. More complex deep learning models such as convolutional neural networks [89] and recurrent neural networks [90] can be applied but with higher computational costs.

It is important to note that, our data, with two main intensive care scoring systems, MODS and NEMS, were collected from two ICUs in London, Ontario, Canada, and the results may not be consistent with those in other ICUs outside of London, Ontario. For future work, a larger data set including the cases in several different ICUs from the Critical Care Information System in Ontario

will be obtained for further analysis and validation based on [91]. COVID-19 patients will also be included in the new data set for predictive modelling.

Table 5: Comparison on AUC values in Mortality prediction between previous works and our study.

Model	AUC value
Logistic regression in [16]	0.760
Logistic regression in [15]	0.767
Logistic regression in our study	0.795
Random forest (RF) in [15]	0.751
Random forest in our study	0.788
Neural network (NN) in [15]	0.638
Neural network (NN) in our study	0.789

Table 6: Comparison on AUC values in LOS prediction between work in [15] and our study.

	Logistic regression	RF	NN
LOS cutting at 7 days in [15]	0.701	0.677	0.606
LOS cutting at 21 days in [15]	0.635	0.622	0.526
LOS cutting at 5 days in our study	0.681	0.689	0.682

CRediT authorship contribution statement

Chengqian Xian: Conception and design of the study, Implementation of statistical analyses, Writing – original draft, Writing – review & editing. **Camila P.E. de Souza:** Conception and design of the study, Writing – review & editing. **Felipe F. Rodrigues** Conception and design of the study, Writing – review & editing.

Acknowledgments

This research work is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). All authors approved the version of the manuscript to be published.

Appendix A Tables

Table A1: Report card based on the ICU data set for mortality and LOS predictions

Problem statement	Predict ICU patient mortality and LOS, respectively, based on patient general characteristics and admission characteristics		
Data gathering	Data collected between January 1st, 2015 and May 31st, 2021, at two teaching hospital ICUs in Ontario, Canada		
Data distribution	15, 350 patients in total; For mortality, 3, 387 of patients deceased and 11, 963 discharged; For length of stay, 10, 939 of patients stayed ≤ 5 days and 4, 411 > 5 days		
Sampling	No sampling		
Data quality	No missing values after data management		
Data preprocessing methods	To build neural networks model, min-max normalization for numeric predictors and one-hot encoding scheme for nominal categorical predictors are conducted		
Feature engineering and vectorizing	No		
Performance estimation			
Parameter optimization	Yes		
	Search space	Random forests	$mtry \in \{1, 2, ..., 23\}$; $ntree \in \{100, 500, 1000\}$
		Neural networks	number of hidden layers $\in \{1, 2\}$; number of neurons on each hidden layer $\in \{1, 2, .., 5\}$
	Search algorithm	Grid search	
Data split	Training set (70%) and validation set (30%); Five-fold validation within the training set for parameter optimization		
Algorithm	Logistic regression, Random forests, Neural networks		
Sampling	No sampling		
Performance metric	AUC from the ROC curve, sensitivity (recall), specificity, accuracy, MCC, PPV (precision), NPV and F1 score		
Performance evaluation	In mortality prediction, logistic regression results in the highest AUC of 0.795; In LOS prediction, random forests results in the highest AUC of 0.689		
Model deployment			
Data used			
Model validity	**Continuous improvement**	No continuous improvement	
	Robustness	No statement about the suitability possible	
Sampling	No sampling		
Algorithm	Logistic regression, Random forests, Neural networks		
	Parameters	Random forests	$mtry = 3$ and $ntree = 1000$
		Neural networks	Mortality prediction: one hidden layer with two neurons; LOS prediction: one hidden layer with one neuron

Note: bold writing indicates a problem characteristic or choice from the report card.

Table A2: Report card based on the ICU data set for LMClass prediction

Problem statement	Predict ICU patient LMClass (LOS-Mortality Class) based on patient general characteristics and admission characteristics		
Data gathering	Data collected between January 1st, 2015 and May 31st, 2021, at two teaching hospital ICUs in Ontario, Canada		
Data distribution	15, 350 patients in total; LMClass: short stay and deceased (2, 275 patients); short stay and discharged (8, 664 patients); long stay (4, 411 patients)		
Sampling	No sampling		
Data quality	No missing values after data management		
Data preprocessing methods	To build neural networks model, min-max normalization for numeric predictors and one-hot encoding scheme for nominal categorical predictors are conducted		
Feature engineering and vectorizing	No		
Performance estimation			
Parameter optimization	Yes		
	Search space	Random forests	$mtry \in \{1, 2, ..., 23\}$; $ntree \in \{100, 500, 1000\}$
		Neural networks	one hidden layer; number of neurons $\in \{1, 2, .., 10\}$
	Search algorithm	Grid search	
Data split	Training set (70%) and validation set (30%); Five-fold validation within the training set for parameter optimization		
Algorithm	Multinomial regression, Random forests, Neural networks		
Sampling	No sampling		
Performance metric	Accuracy, balanced accuracy, Kappa statistic		
Performance evaluation	Multinomial regression results in the highest accuracy of 0.63; Neural networks model results in the highest balanced accuracy of 0.5 and Kappa statistic of 0.303		
Model deployment			
Data used			
Model validity	**Continuous improvement**	No continuous improvement	
	Robustness	No statement about the suitability possible	
Sampling	No sampling		
Algorithm	Multinomial regression, Random forests, Neural networks		
	Parameters	Random forests	$mtry = 9$ and $ntree = 500$
		Neural networks	one hidden layer with three neurons

Note: bold writing indicates a problem characteristic or choice from the report card.

Table A3: Selected predictors in logistic regression for mortality prediction.

Predictor	Coefficient	Odds ratio [95% CI]	LRT p-value[1]
Haematologic			< 0.0001
None (reference)	-	-	
Minimal	0.38	1.46 [1.25, 1.71]	
Mild	0.45	1.57 [1.25, 1.96]	
Moderate	0.85	2.35 [1.82, 3.02]	
Severe	1.22	3.40 [2.22, 5.20]	
Hepatic			< 0.0001
None (reference)	-	-	
Minimal	0.13	1.14 [0.94, 1.39]	
Mild	0.27	1.30 [0.94, 1.80]	
Moderate	0.87	2.39 [1.57, 3.61]	
Severe	1.20	3.32 [2.14, 5.15]	
Renal			< 0.0001
None (reference)	-	-	
Minimal	0.30	1.36 [1.20, 1.53]	
Mild	0.49	1.64 [1.38, 1.95]	
Moderate	0.57	1.77 [1.37, 2.27]	
Severe	0.13	1.14 [0.86, 1.51]	
Cardiovascular			< 0.0001
None (reference)	-	-	
Minimal	0.27	1.31 [1.16, 1.47]	
Mild	0.57	1.77 [1.44, 2.17]	
Moderate	0.53	1.70 [1.24, 2.32]	
Severe	0.71	2.03 [1.22, 3.42]	
Neurologic			< 0.0001
None (reference)	-	-	
Minimal	0.28	1.32 [1.09, 1.61]	
Mild	0.13	1.14 [0.93, 1.40]	
Moderate	0.06	1.06 [0.84, 1.34]	
Severe	0.89	2.45 [2.13, 2.81]	
Respiratory			< 0.0001
None (reference)	-	-	
Minimal	-0.07	0.94 [0.80, 1.09]	
Mild	0.13	1.14 [0.98, 1.33]	
Moderate	0.33	1.39 [1.20, 1.60]	
Severe	0.52	1.68 [1.37, 2.05]	

[1] p-value of the likelihood ratio test for significance of the corresponding predictor

Table A3 continued: Selected predictors in logistic regression for mortality prediction.

Predictor	Coefficient	Odds ratio [95% CI]	LRT p-value[1]
NEMS Score	0.07	1.07 [1.06, 1.08]	< 0.0001
Basic Monitoring			0.005
No (reference)	-		
Yes	-2.23	0.11 [0.03, 0.48]	
Intracranial Pressure Monitor			< 0.0001
No (reference)	-	-	
Yes	0.74	2.09 [1.46, 2.95]	
Dialysis			0.002
No (reference)	-	-	
Yes	-0.44	0.65 [0.49, 0.85]	
Intra-Aortic Balloon Pump			0.119
No (reference)	-	-	
Yes	-0.40	0.67 [0.40, 1.11]	
Other Interventions			0.001
No (reference)	-	-	
Yes	-0.21	0.81 [0.72, 0.92]	
Interventions Outside this Unit			< 0.0001
No (reference)	-	-	
Yes	-0.43	0.65 [0.57, 0.75]	
Age	0.03	1.03 [1.03, 1.04]	< 0.0001
Sex			0.001
Female (reference)	-	-	
Male	-0.17	0.84 [0.76, 0.94]	
Patient Category			0.001
Medical (reference)	-	-	
Surgical	-0.25	0.78 [0.67, 0.91]	
Admission Source			< 0.0001
Emergency (reference)	-	-	
Operating Room	-0.94	0.39 [0.31, 0.48]	
Outside Hospital/Other	-0.01	0.99 [0.86, 1.13]	
Unit/Ward/Stepdown	0.36	1.43 [1.23, 1.66]	
Admission Diagnosis			< 0.0001
Cardiovascular (reference)	-	-	
Gastrointestinal	-0.45	0.64 [0.51, 0.80]	
Neurological	-0.06	0.94 [0.79, 1.14]	
Other	-0.58	0.56 [0.47, 0.66]	
Respiratory	-0.48	0.62 [0.53, 0.72]	
Trauma	-0.11	0.89 [0.68, 1.16]	

[1] p-value of the likelihood ratio test for significance of the corresponding predictor

Table A4: Selected predictors in logistic regression for LOS prediction.

Predictor	Coefficient	Odds ratio [95% CI]	LRT p-value[1]
MODS Score	0.04	1.04 [1.02, 1.06]	0.012
Cardiovascular			0.052
None (reference)	-	-	
Minimal	0.09	1.09 [0.98, 1.21]	
Mild	-0.19	0.83 [0.68, 1.00]	
Moderate	-0.22	0.81 [0.59, 1.09]	
Severe	-0.29	0.75 [0.45, 1.22]	
Respiratory			< 0.0001
None (reference)	-	-	
Minimal	0.12	1.12 [0.98, 1.28]	
Mild	0.29	1.34 [1.16, 1.53]	
Moderate	0.40	1.50 [1.30, 1.73]	
Severe	0.17	1.19 [0.97, 1.46]	
NEMS Score	0.04	1.04 [1.04, 1.05]	< 0.0001
Central Venous Line			< 0.0001
No (reference)	-	-	
Yes	0.46	1.59 [1.41, 1.79]	
Arterial Line			0.041
No (reference)	-	-	
Yes	0.23	1.25 [1.12, 1.41]	
Intracranial Pressure Monitor			0.001
No (reference)	-	-	
Yes	0.67	1.96 [1.43, 2.66]	
Dialysis			0.001
No (reference)	-	-	
Yes	-0.22	0.80 [0.64, 0.99]	
Interventions Outside this Unit			0.007
No (reference)	-	-	
Yes	-0.13	0.88 [0.78, 0.98]	

[1] p-value of the likelihood ratio test for significance of the corresponding predictor

Table A4 continued: Selected predictors in logistic regression for LOS prediction.

Predictor	Coefficient	Odds ratio [95% CI]	LRT p-value[1]
Sex			0.026
Female (reference)	-	-	
Male	0.11	1.12 [1.02, 1.23]	
Admission Source			< 0.0001
Emergency (reference)	-	-	
Operating Room	-0.33	0.72 [0.62, 0.83]	
Outside Hospital/Other	0.47	1.61 [1.43, 1.81]	
Unit/Ward/Stepdown	0.34	1.40 [1.23, 1.60]	
Admission Diagnosis			< 0.0001
Cardiovascular (reference)	-	-	
Gastrointestinal	0.21	1.23 [1.01, 1.49]	
Neurological	0.23	1.25 [1.06, 1.49]	
Other	0.18	1.20 [1.03, 1.39]	
Respiratory	0.55	1.73 [1.51, 1.99]	
Trauma	0.86	2.37 [1.94, 2.90]	

[1] p-value of the likelihood ratio test for significance of the corresponding predictor

Table A5: Selected predictors in multinomial regression for LMClass prediction.

Predictor	short & deceased		long stay	
	Odds Ratio	95% CI	Odds Ratio	95% CI
MODS Score	1.40	[1.34, 1.46]	1.15	[1.11, 1.19]
Renal				
None (reference)	-	-	-	-
Minimal	1.03	[0.87, 1.21]	0.99	[0.87, 1.13]
Mild	0.99	[0.78, 1.26]	1.02	[0.84, 1.25]
Moderate	0.83	[0.59, 1.18]	1.21	[0.92, 1.60]
Severe	0.30	[0.21, 0.45]	0.69	[0.52, 0.92]
Neurologic				
None (reference)	-	-	-	-
Minimal	1.06	[0.83, 1.36]	0.97	[0.81, 1.17]
Mild	0.54	[0.41, 0.72]	0.98	[0.81, 1.18]
Moderate	0.40	[0.29, 0.55]	0.79	[0.64, 0.99]
Severe	0.85	[0.68, 1.07]	0.95	[0.79, 1.14]
Respiratory				
None (reference)	-	-	-	-
Minimal	0.67	[0.56, 0.81]	0.96	[0.84, 1.11]
Mild	0.61	[0.49, 0.75]	1.11	[0.95, 1.30]
Moderate	0.60	[0.48, 0.75]	1.28	[1.08, 1.52]
Severe	0.60	[0.45, 0.82]	1.10	[0.85, 1.42]
NEMS Score	1.11	[1.10, 1.13]	1.07	[1.06, 1.08]
Basic Monitoring				
No (reference)	-	-	-	-
Yes	0.07	[0.01, 0.35]	0.70	[0.08, 5.94]
Central Venous Line				
No (reference)	-	-	-	-
Yes	1.04	[0.87, 1.23]	1.54	[1.36, 1.74]
Arterial Line				
No (reference)	-	-	-	-
Yes	0.94	[0.80, 1.11]	1.22	[1.08, 1.38]
Intracranial Pressure Monitor				
No (reference)	-	-	-	-
Yes	2.05	[1.29, 3.25]	2.34	[1.65, 3.33]
Dialysis				
No (reference)	-	-	-	-
Yes	0.58	[0.41, 0.82]	0.70	[0.54, 092]
Other Interventions Within Unit				
No (reference)	-	-	-	-
Yes	0.74	[0.64, 0.86]	0.94	[0.84, 1.05]
Interventions Outside Unit				
No (reference)	-	-	-	-
Yes	0.56	[0.47, 0.66]	0.76	[0.67, 0.86]

Table A5 continued: Selected predictors in multinomial regression for LMClass prediction.

Predictor	short & deceased		long stay	
	Odds Ratio	95% CI	Odds Ratio	95% CI
Age	1.03	[1.03, 1.04]	1.01	[1.00, 1.01]
Sex				
Female (reference)	-	-	-	-
Male	0.76	[0.67, 0.86]	1.04	[0.94, 1.14]
Admission Source				
Emergency (reference)	-	-	-	-
Operating Room	0.27	[0.21, 0.33]	0.53	[0.46, 0.62]
Outside Hospital/Other	1.03	[0.87, 1.22]	1.63	[1.43, 1.85]
Unit/Ward/Stepdown	1.50	[1.26, 1.80]	1.58	[1.37, 1.83]
Admission Diagnosis				
Cardiovascular (reference)	-	-	-	-
Gastrointestinal	0.58	[0.44, 0.77]	1.00	[0.81, 1.23]
Neurological	1.19	[0.95, 1.48]	1.21	[1.00, 1.46]
Other	0.63	[0.52, 0.77]	1.00	[0.85, 1.18]
Respiratory	0.66	[0.54, 0.79]	1.45	[1.24, 1.69]
Trauma	1.24	[0.91, 1.67]	2.28	[1.83, 2.85]

Table A6: MODS Components (Adapted From Marshal et al 1995[10]).

Organ System	Indicator of Dysfunction	Degree of Dysfunction				
		None (0)	Minimal (1)	Mild (2)	Moderate (3)	Severe (4)
Respiratory	PaO_2/F_IO_2	> 300	$226 - 300$	$151 - 225$	$76 - 150$	≤ 75
Renal	Creatinine $(mmol/L)$	≤ 100	$101 - 200$	$201 - 350$	$351 - 500$	> 500
Cardiovascular	Pressure-adjusted rate	≤ 10.0	$10.1 - 15.0$	$15.1 - 20.0$	$20.1 - 30.0$	> 30.0
Hematological	Platelets $(\times 10^3/mm^3)$	> 120	$80 - 120$	$50 - 80$	$20 - 50$	< 20
Hepatic	Bilirubin $(\mu mol/L)$	≤ 20	$21 - 60$	$61 - 120$	$121 - 240$	> 240
Neurological	Glasgow Coma Score	15	$13 - 14$	$10 - 12$	$7 - 9$	≤ 6

Table A7: NEMS Components (Adapted From Miranda et al 1997 [12]).

Item	Points
1. *Basic monitoring*: hourly vital signs, regular record and calculation of fluid balance	9
2. *Intravenous medication*: bolus or continuously, not including vasoactive drugs	6
3. *Mechanical ventilatory support*: any form of mechanical/assisted ventilation, with or without PEEP (e. g., continuous positive airway pressure), with or without muscle relaxants	12
4. *Supplementary ventilatory care*: breathing spontaneously through endotracheal tube; supplementary oxygen any method, except if (3) applies	3
5. *Single vasoactive medication*: any vasoactive drug	7
6. *Multiple vasoactive medication*: more than one vasoactive drug, regardless of type and dose	12
7. *Dialysis techniques*: all	6
8. *Specific interventions in the ICU*: such as endotracheal intubation, introduction of pacemaker, cardioversion, endoscopy, emergency operation in the past 24 h, gastric lavage; routine interventions such as X-rays, echocardiography, electrocardiography, dressings, introduction of venous or arterial lines, are not included	5
9. *Specific interventions outside the ICU*: such as surgical intervention or diagnostic procedure; the intervention/procedure is related to the severity of illness of the patient and makes an extra demand upon manpower efforts in the ICU	6
Total	56

Appendix B Figures

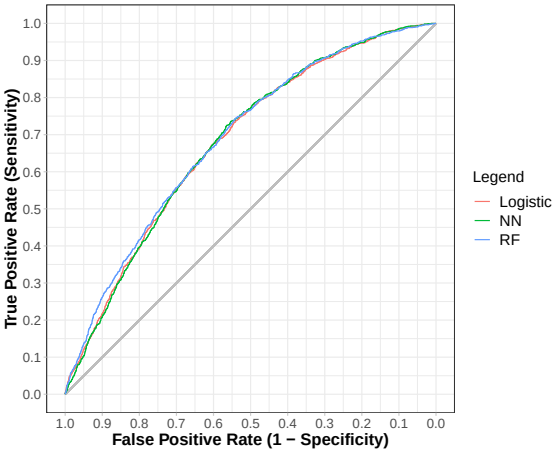

Fig. B1: ROC curves for LOS prediction in validation set.

Red: logistic regression; Green: random forests; Blue: neural networks

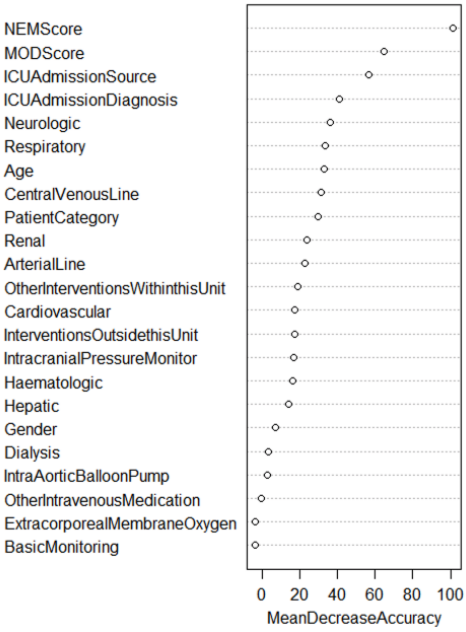

Fig. B2: Importance of predictors based on RF model in LMClass prediction.

References

[1] Salluh J, Chiche JD, Reis CE. New perspectives to improve critical care benchmarking. Ann Intensive Care. 2018;8(1):17.

[2] Silva Ramos F, Figueira Salluh JI. Data-driven management for intensive care units. ICU Management & Practice. 2019;19.

[3] Rapsang AG, Shyam DC. Scoring systems in the intensive care unit: A compendium. Indian journal of critical care medicine. 2014;18(4):220–228.

[4] Salluh J, Soares M. ICU severity of illness scores: APACHE, SAPS and MPM. Current Opinion in Critical Care. 2014;20(5):557–565.

[5] Le Gall JR, A N, F H, et al. Mortality prediction using SAPS II: an update for French intensive care units. Crit Care. 2005;9:R645.

[6] Rubenfeld G, Angus D, Pinsky M, Curtis J, Connors AJ, GR B. Outcomes research in critical care: Results of the American Thoracic Society Critical Care Assembly Workshop on Outcomes Research. Am J Respir Crit Care Med. 1999;160:358–367.

[7] 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) $IV^{\star}$. Critical Care Medicine. 2006;34(10):2517–2529.

[8] Le GJ, Lemeshow S, Saulnier F. A New Simplified Acute Physiology Score (SAPS II) Based on a European/North American Multicenter Study. JAMA. 1993;270(24):2957–2963.

[9] Lemeshow S, Teres D, Klar J, Avrunin J, Gehlbach S, J R. Probability Models (MPM II) Based on an International Cohort of Intensive Care Unit Patients. JAMA. 1993;270(20):2478–2486.

[10] Marshall JC, Cook DJ, Christou NV, R BG, Sprung CL, Sibbald WJ. Multiple organ dysfunction score: a reliable descriptor of a complex clinical outcome. Critical Care Medicine. 1995;23(10):1638–1652.

[11] Vincent J, de Mendonça A, Cantraine F, Moreno R, Takala J, Suter P, 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. Working group on "sepsis-related problem" of the European Society of Intensive Care Medicine. Critical care medicine. 1998;26(11):1793–800.

[12] Reis MD, Moreno R, Iapichino G. Nine equivalents of nursing manpower use score (NEMS). Intensive Care Medicine. 1997;23(7):760–765.

[13] Cullen D, Civetta J, Briggs B, Ferrara L. Therapeutic intervention scoring system: a method for quantitative comparison of patient care. Crit Care Med. 1974;2(2):57–60.

[14] Rothen HU, Küng V, Ryser DH, Zürcher R, Regli B. Validation of "nine equivalents of nursing manpower use score" on an independent data sample. Intensive Care Medicine. 1999;25(6):606–611.

[15] Rodrigues FF. Three Essays on Intensive Care Unit Capacity Planning. Electronic Thesis and Dissertation Repository. 2018;5984.

[16] Kao R, Priestap F, Donner A. To develop a regional ICU mortality prediction model during the first 24 h of ICU admission utilizing MODS and NEMS with six other independent variables from the Critical Care Information System (CCIS) Ontario, Canada. Journal of intensive care. 2016;4(1):1–12.

[17] Verburg IWM, Atashi A, Eslami S, Holman R, Abu-Hanna A, de Jonge E, et al. Which models can I use to predict adult ICU length of stay? A systematic review. Critical care medicine. 2017;45(2):e222–e231.

[18] Faraway J. Extending the Linear Model with R. New York: Chapman and Hall/CRC; 1981. Available from: https://doi.org/10.1201/9781315382722.

[19] Terza J. Estimating endogenous treatment effects in retrospective data analysis. Value in Health. 1999;2(6):429–434.

[20] Moran JL, Solomon PJ, for Outcome AC, of the Australian REC, (ANZICS) NZICS. A review of statistical estimators for risk-adjusted length of stay: analysis of the Australian and New Zealand intensive care adult patient data-base, 2008–2009. BMC medical research methodology. 2012;12:1–17.

[21] Lingsma HF, Bottle A, Middleton S, Kievit J, Steyerberg EW, Marang-Van De Mheen PJ. Evaluation of hospital outcomes: the relation between length-of-stay, readmission, and mortality in a large international administrative database. BMC health services research. 2018;18(1):1–10.

[22] Rush B, Celi LA, Stone DJ. Applying machine learning to continuously monitored physiological data. Journal of Clinical Monitoring and Computing. 2019;33(5):887–893. https://doi.org/10.1007/s10877-018-0219-z.

[23] Vellido A, Ribas V, Morales C, et al. Machine learning in critical care: state-of-the-art and a sepsis case study. Biomed Eng Online. 2018;17(1):135.

[24] Xia H, Daley BJ, Petrie A, Zhao X. A neural network model for mortality prediction in ICU. In: 2012 Computing in Cardiology; 2012. p. 261–264.

[25] Asteris PG, Gavriilaki E, Touloumenidou T, Koravou EE, Koutra M, Papayanni P, et al. Genetic prediction of ICU hospitalization and mortality in COVID-19 patients using artificial neural networks. Journal of Cellular and Molecular Medicine. 2022;26(5):1445–1455. https://doi.org/https://doi.org/10.1111/jcmm.17098.

[26] Fusaro MV, Becker C, Scurlock C. Evaluating tele-ICU implementation based on observed and predicted ICU mortality: a systematic review and meta-analysis. Critical care medicine. 2019;47(4):501–507.

[27] Keuning BE, Kaufmann T, Wiersema R, Granholm A, Pettilä V, Møller MH, et al. Mortality prediction models in the adult critically ill: A scoping review. Acta Anaesthesiologica Scandinavica. 2020;64(4):424–442.

[28] John LM, Peter B, Patricia JS, Carol G, Graeme KH. Mortality and length-of-stay outcomes, 1993-2003, in the binational Australian and New Zealand intensive care adult patient database. Critical Care Medicine. 2008;p. 46—61.

[29] Zampieri FG, et al. Customization and external validation of the Simplified Mortality Score for the Intensive Care Unit (SMS-ICU) in Brazilian critically ill patients. Journal of Critical Care. 2020;59:94–100.

[30] Lemeshow S, Teres D, Avrunin JS, Gage RW. Refining intensive care unit outcome prediction by using changing probabilities of mortality. Critical Care Medicine. 1988;16(5):470–477.

[31] Niskanen M, Reinikainen M, Pettilä V. Case-mix-adjusted length of stay and mortality in 23 Finnish ICUs. Intensive Care Med. 2009;35:1060–1067.

[32] Engelhardt LJ, Balzer F, Müller MC, Grunow JJ, Spies CD, Christopher KB, et al. Association between potassium concentrations, variability and supplementation, and in-hospital mortality in ICU patients: a retrospective analysis. Annals of intensive care. 2019;9:1–11.

[33] Zhao Z, Chen A, Hou W, Graham JM, Li H, Richman PS, et al. Prediction model and risk scores of ICU admission and mortality in COVID-19. PloS one. 2020;15(7):e0236618.

[34] Wilcox ME, Harrison DA, Patel A, Rowan KM. Higher ICU capacity strain is associated with increased acute mortality in closed ICUs. Critical Care Medicine. 2020;48(5):709–716.

[35] Ahlström B, Frithiof R, Hultström M, Larsson IM, Strandberg G, Lipcsey M. The swedish covid-19 intensive care cohort: Risk factors of ICU admission and ICU mortality. Acta Anaesthesiologica Scandinavica. 2021;65(4):525–533.

[36] Smail SW, Babaei E, Amin K. Hematological, Inflammatory, Coagulation, and Oxidative/Antioxidant Biomarkers as Predictors for Severity and Mortality in COVID-19: A Prospective Cohort-Study. International Journal of General Medicine. 2023;p. 565–580.

[37] Lavrentieva A, Kaimakamis E, Voutsas V, Bitzani M. An observational study on factors associated with ICU mortality in Covid-19 patients and critical review of the literature. Scientific Reports. 2023;13(1):7804.

[38] Xie Z, Hong YR, Tanner R, Marlow NM. People with functional disability and access to health care during the COVID-19 pandemic: a US population-based study. Medical care. 2023;61(2):58–66.

[39] Yaseliani M, Khedmati M. Prediction of Heart Diseases Using Logistic Regression and Likelihood Ratios. International Journal of Industrial Engineering & Production Research. 2023;34(1):1–15.

[40] Norrie J. Mortality prediction in ICU: a methodological advance. Lancet Respir Med. 2015;3(1):5–6.

[41] Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. Lancet Respir Med. 2015;3(1):42–52.

[42] Ahmed FS, Ali L, Joseph BA, FACS; Ikram A, Ul Mustafa R, Bukhari SAC. A statistically rigorous deep neural network approach to predict mortality in trauma patients admitted to the intensive care unit. Journal of Trauma and Acute Care Surgery. 2020;89(4):736–742.

[43] Iwase S, Nakada Ta, Shimada T, Oami T, Shimazui T, Takahashi N, et al. Prediction algorithm for ICU mortality and length of stay using machine learning. Scientific Reports. 2022;12(1):12912.

[44] Luo C, Zhu Y, Zhu Z, Li R, Chen G, Wang Z. A machine learning-based risk stratification tool for in-hospital mortality of intensive care unit patients with heart failure. Journal of Translational Medicine. 2022;20(1):136.

[45] Elhazmi A, Al-Omari A, Sallam H, Mufti HN, Rabie AA, Alshahrani M, et al. Machine learning decision tree algorithm role for predicting mortality in critically ill adult COVID-19 patients admitted to the ICU.

Journal of Infection and Public Health. 2022;15(7):826–834.

[46] Jamshidi E, Asgary A, Tavakoli N, Zali A, Setareh S, Esmaily H, et al. Using machine learning to predict mortality for COVID-19 patients on day 0 in the ICU. Frontiers in Digital Health. 2022;3:210.

[47] ;.

[48] Kuno T, Sahashi Y, Kawahito S, Takahashi M, Iwagami M, Egorova NN. Prediction of in-hospital mortality with machine learning for COVID-19 patients treated with steroid and remdesivir. Journal of Medical Virology. 2022;94(3):958–964.

[49] Baker TB, Loh WY, Piasecki TM, Bolt DM, Smith SS, Slutske WS, et al. A machine learning analysis of correlates of mortality among patients hospitalized with COVID-19. Scientific Reports. 2023;13(1):4080.

[50] Awad A, Bader-El-Den M, McNicholas J. Patient length of stay and mortality prediction: a survey. Health services management research. 2017;30(2):105–120.

[51] Peres IT, Hamacher S, Oliveira FLC, Thomé AMT, Bozza FA. What factors predict length of stay in the intensive care unit? Systematic review and meta-analysis. Journal of Critical Care. 2020;60:183–194.

[52] Peres IT, Hamacher S, Oliveira FLC, Bozza FA, Salluh JIF. Prediction of intensive care units length of stay: a concise review. Revista Brasileira de Terapia Intensiva. 2021;33:183–187.

[53] 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.

[54] Van Houdenhoven M, Nguyen DT, Eijkemans MJ, Steyerberg EW, Tilanus HW, Gommers D, et al. Optimizing intensive care capacity using individual length-of-stay prediction models. Critical Care. 2007;11(2):1–10.

[55] Kramer AA, Zimmerman JE. The relationship between hospital and intensive care unit length of stay. Critical care medicine. 2011;39(5):1015–1022.

[56] Houthooft R, Ruyssinck J, van der Herten J, Stijven S, Couckuyt I, Gadeyne B, et al. Predictive modelling of survival and length of stay in critically ill patients using sequential organ failure scores. Artificial Intelligence in Medicine. 2015;p. 191–207.

[57] Peres IT, Hamacher S, Oliveira FLC, Bozza FA, Salluh JIF. Data-driven methodology to predict the ICU length of stay: A multicentre study of 99,492 admissions in 109 Brazilian units. Anaesthesia Critical Care & Pain Medicine. 2022;41(6):101142.

[58] Alsinglawi B, Alnajjar F, Mubin O, Novoa M, Alorjani M, Karajeh O, et al. Predicting length of stay for cardiovascular hospitalizations in the intensive care unit: Machine learning approach. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE; 2020. p. 5442–5445.

[59] Mekhaldi RN, Caulier P, Chaabane S, Chraibi A, Piechowiak S. Using machine learning models to predict the length of stay in a hospital setting. In: Trends and Innovations in Information Systems and Technologies: Volume 1. Springer; 2020. p. 202–211.

[60] Tu JV G. Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery. Comput Biomed Res. 1992;26:666–672.

[61] Alghatani K, Ammar N, Rezgui A, Shaban-Nejad A, et al. Predicting intensive care unit length of stay and mortality using patient vital signs: machine learning model development and validation. JMIR medical informatics. 2021;9(5):e21347.

[62] Shryane N, Pampaka M, Aparicio-Castro A, Ahmad S, Elliot MJ, Kim J, et al. Length of stay in icu of covid-19 patients in england, march-may 2020. International Journal of Population Data Science. 2020;5(4).

[63] Vekaria B, Overton C, Wiśniowski A, Ahmad S, Aparicio-Castro A, Curran-Sebastian J, et al. Hospital length of stay for COVID-19 patients: Data-driven methods for forward planning. BMC Infectious Diseases. 2021;21(1):1–15.

[64] Agarwal N, Biswas B, Singh C, Nair R, Mounica G, Jha AR, et al. Early Determinants of Length of Hospital Stay: A Case Control Survival Analysis among COVID-19 Patients admitted in a Tertiary Healthcare Facility of East India. Journal of Primary Care & Community Health. 2021;12:21501327211054281.

[65] Kühl N, Hirt R, Baier L, Schmitz B, Satzger G. How to conduct rigorous supervised machine learning in information systems research: the supervised machine learning report card. Communications of the Association for Information Systems. 2021;48(1):46.

[66] Ho TK. The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence.

1998;20(8):832–844.

[67] Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18–22.

[68] Breiman L. Random Forests. Machine Learning. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324.

[69] Riedmiller MA. Rprop - Description and Implementation Details; 1994. .

[70] Harris D, Harris S. Digital Design and Computer Architecture 2nd Edition. San Francisco, Calif.: Morgan Kaufmann; 2012.

[71] McHugh ML. Interrater reliability: the kappa statistic. Biochemia medica. 2012;22(3):276–282.

[72] Nahm FS. Receiver operating characteristic curve: overview and practical use for clinicians. Korean journal of anesthesiology. 2022;75(1):25–36.

[73] Abdul Bujang SD, Fujita H, et al. Imbalanced Classification Methods for Student Grade Prediction: A Systematic Literature Review. 2022;.

[74] Laupland KB, Kirkpatrick AW, Kortbeek JB, Zuege DJ. Long-term mortality outcome associated with prolonged admission to the ICU. chest. 2006;129(4):954–959.

[75] Taccone P, Langer T, Grasselli G.: Do we really need postoperative ICU management after elective surgery? No, not any more! Springer.

[76] Kumar P, Bhatnagar R, Gaur K, Bhatnagar A. Classification of imbalanced data: review of methods and applications. In: IOP conference series: materials science and engineering. vol. 1099. IOP Publishing; 2021. p. 012077.

[77] Morton A, Marzban E, Giannoulis G, Patel A, Aparasu R, Kakadiaris IA. A comparison of supervised machine learning techniques for predicting short-term in-hospital length of stay among diabetic patients. In: 2014 13th International Conference on Machine Learning and Applications. IEEE; 2014. p. 428–431.

[78] 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 medical informatics and decision making. 2010;10(1):1–16.

[79] Livieris IE, Kotsilieris T, Dimopoulos I, Pintelas P. Decision support software for forecasting patient's length of stay. Algorithms. 2018;11(12):199.

[80] Hassan A, Anderson C, Kypson A, Kindell L, Ferguson TB, Chitwood Jr WR, et al. Clinical outcomes in patients with prolonged intensive care unit length of stay after cardiac surgical procedures. The Annals of thoracic surgery. 2012;93(2):565–569.

[81] Hermans G, Van Aerde N, Meersseman P, Van Mechelen H, Debaveye Y, Wilmer A, et al. Five-year mortality and morbidity impact of prolonged versus brief ICU stay: a propensity score matched cohort study. Thorax. 2019;74(11):1037–1045.

[82] Soares M, Salluh JI, Torres VB, Leal JV, Spector N. Short-and long-term outcomes of critically ill patients with cancer and prolonged ICU length of stay. Chest. 2008;134(3):520–526.

[83] Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32–35.

[84] Goss EP, Ramchandani H. Survival prediction in the intensive care unit: a comparison of neural networks and binary logit regression11Study supported by a grant from Bishop-Clarkson Hospital and Applied Information Management Institute. Socio-Economic Planning Sciences. 1998;32(3):189–198. https://doi.org/https://doi.org/10.1016/S0038-0121(97)00039-6.

[85] Trigg H. An investigation of methods to enhance the performance of artificial neural networks used to estimate medical outcomes. University of New Brunswick, Electrical Engineering Department; 1997.

[86] Frize M, Ennett CM, Stevenson M, Trigg HCE. Clinical decision support systems for intensive care units: using artificial neural networks. Medical Engineering Physics. 2001;23(3):217–225. https://doi.org/https://doi.org/10.1016/S1350-4533(01)00041-8.

[87] Schmidt-Hieber J. Nonparametric regression using deep neural networks with ReLU activation function. The Annals of Statistics. 2020;48(4):1875 – 1897. https://doi.org/10.1214/19-AOS1875.

[88] Wu H, Fan Y, Lv J. Statistical insights into deep neural network learning in subspace classification. Stat. 2020;9(1):e273. https://doi.org/https://doi.org/10.1002/sta4.273.

[89] Fukushima K, Miyake S. Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Visual Pattern Recognition. In: Amari Si, Arbib MA, editors. Competition and Cooperation in Neural Nets. Berlin, Heidelberg: Springer Berlin Heidelberg; 1982. p. 267–285.

[90] Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of

Sciences. 1982;79(8):2554–2558. https://doi.org/10.1073/pnas.79.8.2554.

[91] Priestap F, Kao R, Martin CM. External validation of a prognostic model for intensive care unit mortality: a retrospective study using the Ontario Critical Care Information System. Can J Anesth/J Can Anesth. 2020;67:981–991.