Informatics in Medicine Unlocked 28 (2022) 100861

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

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

The prediction power of machine learning on estimating the sepsis mortality in the intensive care unit

Mehtap Selcuk [a], Oguz Koc [b], A. Sevtap Kestel [c,*]

[a] Reanimation and Intensive Care Department, Acibadem University, Kerem Aydinlar Kampusü, Kayisdagi Cad. No:32, Ataşehir, 34758, Istanbul, Turkey
[b] Department of Financial Mathematics, Institute of Applied Mathematics, Middle East Technical University, Universiteler, Dumlupinar Bulvari No:1, 06800, Cankaya, Ankara, Turkey
[c] Department of Actuarial Sciences, Institute of Applied Mathematics and Biomedical Engineering, Graduate School of Natural Sciences, Middle East Technical University, Universiteler, Dumlupinar Bulvari No:1, 06800, Cankaya, Ankara, Turkey

ARTICLE INFO

Keywords:
APACHE II
Machine learning
Generated ensemble
SAPS II
Sepsis mortality
SOFA

ABSTRACT

Background: The prediction of sepsis mortality of intensive care unit (ICU) observations using machine learning (ML) methods is hypothesized to yield better or as good as performance compared to the prognostic scores. This paper aims to show that the accuracy of ML in sepsis mortality estimation can be superior and supportive knowledge to SAPS II, APACHE II, and SOFA (traditional) scores even under small sample restrictions.
Methods: The retrospective collection of data from the patients (n = 200) admitted to ICU of Acibadem Hospital, Istanbul-Turkey, between 2015 and 2020 is utilized to detect the sepsis mortality risk using eight ML methods and a generated ensemble model along with the traditional prognostic scores. The mortality as a decisive indicator is evaluated according to the explanatory variables included quantifying the traditional scores. In the calibration of the data, five different predetermined splits of the random samples are used for the training and the validation of the ML methods. The efficiency of the prediction results of ML methods and the traditional scoring methods are investigated by AUC-ROC curves and other accuracy indicators. Consecutive processes of Box-Cox and Min-Max transformations on data and parameter optimization are performed to increase the efficiency of algorithms.
Results: The accuracy in the mortality prediction is achieved the best by the Multi-Layer Perceptron algorithm compared to SAPS II and APACHE II methods and is as good as the one with what SOFA predicts. The prediction power of the best performing ML methods for APACHE II, SAPS II, and SOFA are found to be 84.45%, 85.25% and 73.47%, respectively. The ensemble of eight ML methods is found to increase the performance around 2% in APACHE II score.
Conclusions: The outcomes of this study have clinical merits in evaluating the potential use of ML methods in predicting ICU mortality superior to traditional scores APACHE II, SAPS II, and as good as SOFA. Additionally, it explores which of the variables contributing to sepsis mortality risk should be taken as apriori information in treating the patients and requires fewer number of explanatory variables, with reliable prediction powers even for considerably small sample size data sets.

1. Introduction

Early and appropriate treatment of sepsis which is a dysregulated immune-mediated host response to infection, would potentially reduce mortality and costs. Mortality due to sepsis among patients who are in the intensive care unit (ICU) depends on many factors, including age,

; *ICU*, Intensive Care Unit; ML, Machine Learning; *APACHE, the Acute Physiology and Chronic Health Evaluation*; *SAPS, Simplified Acute Physiology Score-Simplified Acute Physiology Score*; *SOFA, Sepsis Related Organ Failure Assessment-Sepsis Linked Organ Disease Assessment*; *AUC, Area Under Curve*; *LR, Logistic Regression*; *NN, Neural Networks*; *GNB, Gaussian Naïve Bayes*; *SVM, Support Vector Machine*; *DT, Decision Tree*; *RF, Random Forest*; *XGB, XGBoost*; *MLP, Multi-layer Perceptron Neural Network*; *KNN*, K-nearest Neighbor; *GS, Grid Search*; *TP, True Positive*; *TN*, True Negative..

* Corresponding author. Department of Actuarial Sciences, Institute of Applied Mathematics, Middle East Technical University, Universiteler, Dumlupinar Bulvari No:1, 06800 Cankaya, Ankara, Turkey.
E-mail addresses: selcuk.mehtap@gmail.com, selcuk.mehtap@acibadem.edu.tr (M. Selcuk), oguzkoc@metu.edu.tr (O. Koc), skestel@metu.edu.tr (A.S. Kestel).

https://doi.org/10.1016/j.imu.2022.100861
Received 18 August 2021; Received in revised form 30 December 2021; Accepted 17 January 2022
Available online 20 January 2022

physiological features, and organ status [1]. Besides its high mortality risk, the cost of sepsis can be a burden as the average duration of stay is much higher than other conditions For this reason, techniques for scoring a patient's level of risk for sepsis mortality are essential aids in implementing necessary precautions during treatment, and in predicting prognosis.

Scoring systems for ICU patients offer advantages overdiagnosis-based methods used in other patient groups because ICU patients may have multiple organ failure whose effect may not be tracked quickly during treatment. Among many ICU mortality scores, the Acute Physiology and Chronic Health Evaluation (APACHE) [2], and the Simplified Acute Physiology Score-Simplified Acute Physiology Score (SAPS) [3] are the most widely used. Each of these are calculated based on the observations on the first day of ICU admission. In contrast, Sepsis Related Organ Failure Assessment-Sepsis Linked Organ Disease Assessment (SOFA) [4] scoring is a repetitive method based on data that is collected sequentially over the first days or throughout the entire ICU stay. Prognostic scoring systems (e.g., APACHE, SAPS) are concerned with predicting mortality, whereas organ dysfunction scores (e.g., SOFA) also include morbidity. These scores are calculated on a web-based interface whose values are entered manually. Performance of such scores is assessed by calibrating the data set and determine its statistical significance. One of the best indicators, area-under-curve receiver-operative curve (AUC-ROC) analysis, reveals the specificity and sensitivity of a scoring method, and is used to estimate mortality [5–11].

Machine learning (ML) techniques aim to determine which factors in a data set are strong determinants of certain behaviors in the real world. Implementation of ML methods varies with respect to the type and structure of the data set to be used. In literature related to ICU patients, reports on ML have been limited mostly to the comparative application of neural networks and logistic regression [6–15]. However, recent studies done in predicting ICU sepsis-3 mortality on a big data base (MIMIC III) indicate that the utilization of ML as next genre is necessary and important [16–23]. Application of deep Neural Network with the help of a collocation method to solve partial differential equations [24] and solving the second-order boundary value problem, by creating a collocation method using Neural Network are the recent studies implementing the numerical methods to in the frame of ML techniques [19]. Comparisons of data mining methods with APACHE II and SAPS [25,26] as well as SOFA [27–29] scores for predicting mortality suggest that ML techniques (neural networks and logistic regression) can help to determine and increase the efficiency of intensive care indicators for decision making. However, sepsis mortality stems from multiple sources of organ failure that are interconnected. There is a need to be able to distinguish and accurately predict sepsis-triggered mortality in ICU patients using other ML methods, combined with required data transformations and parameter optimization.

Traditional methods require the manual entry of each values of around 20 variables to predict the mortality risk and each of these methods need different inputs having some in common. Besides, their web-based interfaces do not allow to investigate the effect of additional variable in these black-box systems. On the other hand, in case of having automated ML algorithm which can be as good as or better than traditional scores, is beneficial in many ways: (i) justifying the accuracy of traditional scores, (ii) allowing improvements and changes in the scoring systems. Having these as our motivation, we aim to investigate whether supervised ML techniques can accurately predict mortality condition in ICU patients with sepsis whose critical states at the very first 24-h after their admission to ICU are usually characterized using these well-known scores. An unsupervised learning approach will not be relevant, as the main indicator of sepsis mortality is a binary variable whose value is expected and shown to be influenced by other variables. We hypothesize that, for this patient group, using ML algorithms as adjunct to conventional scores to estimate the ICU mortality as a supporting indicator to conventional scores is reliable and accurate, even for small observation sets. In the content of this study, we compare the predictive power of eight selected ML methods with respect to their hyper-parameters and prognostic scores as well as taking into account an ensemble of the ML algorithms.

To justify the targeted hypothesis, a retrospective data set of 200 patients, registered to ICU unit of Acibadem Hospital, Istanbul-Turkey due to different causes, is used as evidence for the comprehensive application of eight selected ML methods whose prediction powers are compared with respect to their hyper-parameters as well as prognostic diagnose scores. To do so, the same variables which are necessary to calculate mentioned traditional scores are taken as variables for running these ML algorithms. In order to increase the robustness in predicting sepsis mortality risk, we generate an ensemble of eight selected ML algorithms. Besides the traditional performance indicator, AUC score, we compare the rates of true positives (TP) and true negatives (TN), which are the essential and necessary indicators in the ML algorithms.

2. Materials and methods

2.1. Patients, data, and software

The Institutional Review Board of Acibadem University approved this retrospective study, which accessed recorded historical observations of patients admitted to the ICU of Acibadem Kadikoy Hospital in Istanbul, Turkey (Protocol Number: ATADEK-2020-04/10). Data were collected from the records of 200 patients who were admitted to the hospital's ICU for different reasons between 2015 and 2020. All patients were adults (>18 years of age) and fulfilled the conditions defined by the Third Consensus Definitions for Sepsis and Septic Shock [30].

As mentioned, eight ML methods for predicting mortality in this patient group are compared. The variables recorded from each patient's data set are those used by the noted traditional scoring methods (i.e., APACHE II, SAPS II, SOFA). This includes 38 variables (features) in total (Table 1). Web-based interfaces (*e.g.* Ref. [31]) allow users to determine the traditional scores which requires the manual entry of the observed values. It should be emphasized that, depicting the efficiency of an ML algorithm over traditional scores necessitates consistency in the selection of the variables. Upon the admission to the ICU, the most intensive time frame for sepsis is known to be the first 24-hrs as it requires the medical actions based on the important indicators. These are monitored continually and as a representative of this critical period, the maximum (or worst) observation is taken as the threshold to lead to sepsis mortality risk. Therefore, in this study, recorded observations for each variable on each patient base constitute the worst observed within the first 24 h after ICU admission.

All eight ML algorithms are run using Python (3.9.6) software and Sklearn-learn libraries. When applying each algorithm, the total data set (i.e., 200 patients) is randomly categorized into training groups and test groups based on the percentage of deaths in the binary dependent variable "mortality".

2.2. Traditional scoring methods and machine learning

Mortality scoring for ICU patients is well-established in the literature, and web-based programs are available to easily generate these scores. APACHE II, SAPS II, and SOFA scores are determined using certain sets of variables, and some of these are shared by the different methods. In APACHE II, a patient is categorized as critical state at score 71 [2], whereas a SAPS II score above 35 indicates the mortality risk >80% [3]. The SOFA score predicts mortality risk at initial ICU admission and in the following hours based on the status of body systems: respiratory, cardiovascular (and specifically coagulation), hepatic, renal, and neurological. A total SOFA score >3 reflects organ failure [4].

The ability to make accurate predictions using ML methods also requires careful implementation of applications such as hyper-parameter optimization, data transformation, and validation structure

Table 1
The variables and their specifications together with descriptive statistics.

Features (units)	Used in Score	Mean (SD)	Min	Max	CoV*%	JB Test p-value
Age (years)	APACHE SAPS	74.6 (12.5)	22.00	100.00	17.00	4.61E-24
Vital Signs						
Glaskow Coma Score	APACHE SAPS, SOFA	10.60 (3.12)	3.00	15.00	29.00	0.0089
Temperature (C)	APACHE SAPS	37.19 (1.21)	34.00	40.00	3.00	**0.3233**
Systolic BP (mmHg)	SAPS	105.21 (31.72)	40.00	201.00	30.00	0.0003
MAP (mmHg)	APACHE SOFA	58.28 (11.80)	34.00	105.00	20.00	4.24E-18
Heart Rate (bpm)	APACHE SAPS	87.87 (8.49)	20.00	158.00	9.60	0.0204
Respiratory Rate (cpm)	APACHE	25.11 (7.08)	11.21	39.08	28.00	0.0083
Urine Output (ml)	SAPS, SOFA	1168.95 (492.76)	0.00	2500.00	42.00	**0.7044**
Arter Blood Gas						
PaO2 (mmHg)	APACHE SAPS, SOFA	95.09 (50.89)	25.30	322.00	54.00	6.71E-43
Arterial PH	APACHE	7.31 (0.13)	6.90	7.59	2.00	
Bicorbanote (Mmol/l)	SAPS	22.72 (6.94)	7.60	39.70	30.00	**0.1989**
$FiO2^{+}$	APACHE SAPS	0.581 (0.15)	0.28	1.00	26.00	0.0030
Mechanical Ventilation	SAPS, SOFA	0.26 (0.44)				
Laboratory Results						
Sodium (mEq/L)	APACHE SAPS	138.33 (42.10)	122.00	153.00	30.00	**0.4722**
Potassium (mEq/L)	APACHE SAPS	4.17 (0.90)	2.49	6.86	22.00	0.0003
Creatinine (Ht %)	APACHE SAPS, SOFA	1.75 (1.19)	0.26	7.30	68.00	4.91E-36
Hematocrit (mmol/l)	APACHE	32.59 (6.29)	20.00	48.70	19.00	
WBC (10^3/mm^3)	APACHE SAPS	14.48 (7.95)	1.05	45.09	55.00	2.56E-25
Platelets (10^3/mm^3)	SOFA	254.73 (125.58)	8 .00	690.00	49.00	
BUN (mg/dl)	SAPS	43.11 (25.01)	7.00	169.00	58.00	4.55E-29
Biliribun (mg/dl)	SAPS, SOFA	0.74 (0.33)	0.31	2.41	45.00	2.75E-76
Others			**ICU Admission reason**			
Vasopressor	SOFA	0.21 (0.4073)	Cerebrovascular event	0.01 (0.0995)		
Chronic Disease	SAPS	0.41 (0.4909)	Lung cancer	0.03 (0.1706)		
Chronic kidney failure	–	0.06 (0.2280)	Aspiration Pneumonia	0.05 (0.2180)		
Severe organ system insufficiency	APACHE	0.89 (0.3129)	Pancreatic cancer	0.01 (0.0705)		
Acute renal failure	APACHE	0.85 (0.3619)	Hepatic coma	0.01 (0.0995)		
			Brain cancer	0.01 (0.0705)		
			Pneumonia	0.50 (0.5000)		
ICU Death (Exitus)	**Dependent variable**	**0.34 (0.4720)**	Inflammation in COPD	0.03 (0.1706)		
			ARDS	0.01 (0.0705)		

*: coefficient of variation; JB: Jargue-Bera.

techniques. The many combinations of implementing these applications may generate a wide variation in trial numbers, efficiency in prediction, and computing time in algorithm executions.

Among commonly used ML methods, we mainly focus on logistic regression, Gaussian Naive Bayes, Support-vector-machine, decision tree, random forest, XGBoost, K-nearest neighbor and Multi-layer perceptron neural network.

Logistic regression (LR) is nonlinear modeling of a binary dependent variable relative to independent variables which is expressed as

$$P(D|X_1, X_2, \ldots, X_n) = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n)}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n)} \quad (1)$$

where $X_i, i = 1, .., n,$ are the independent and D is the binary dependent variables. The central point in this transformation is that it is designed to identify a probability that is always a number in the range between 0 and 1 due to the S-shape of the logistic function. The classification of an instance is done by a threshold value.

Gaussian Naive Bayes (GNB) enables the ML process to be transparent and clear through posterior distributions, that assign the most probable class to a specified instance defined by its attributes. As a statistical approach it has a simple, but effective methodology which makes the user clear understanding of how it behaves. By considering the class memberships, posterior probabilities of samples are calculated, and the observations are classified to the most probable class. The conditional probability of an observation is calculated as

$$P(C_c|E) = \frac{\mathrm{P}(E|\mathrm{C_c})\mathrm{P}(\mathrm{C_c})}{\sum_{k=1}^{N} P(E|C_k)P(C_k)} \quad (2)$$

where E=($x_1, x_2, \ldots, x_n$) is an observation in the dataset, C_c is the class c. With the help of this probability, an instance is classified by using a threshold. In order to calculate probabilities, $P(E|C_c)$, Gaussian transformation with the mean and the standard deviation of class c is employed.

Support vector machine (SVM) creates a decision boundary to separate the tuples of one class from others, and decides on the features using support vectors from the training points. It searches for an optimal hyperplane (a decision boundary) in the data space to separate the observations into classes using kernel functions which convert the data to a higher dimension. In this higher dimensional space, the algorithm locates support vectors that separate the classes with the highest margins. A decision boundary with weight vector, $W = (w_1, w_2, \ldots, w_n)$, is defined as

$$WX + b = 0 \quad (3)$$

Here, n is the number of features, b is a scaler or bias. The decision is made by comparing an instance is whether above the hyperplane or not. When the instance is above the decision boundary the given equation is positive otherwise, it is negative.

Decision tree (DT) represents nodes as the test on a feature, leaves as class labels, and branches the outcomes. This approach consists of internal nodes, branches, and leaf nodes with a tree structure which is constructed by the entropy or Gini index. These functions measure the impurity level of the nodes, and the branching aims to decrease the impurity level in each step. When a node is pure enough, the branching is stopped, and that node becomes a leaf. Each leaf has a label, and when an observation drops into a specific one the instance takes the label of the leaf. The entropy formula is given as

$$\mathrm{info(S)} = \sum_{i=1}^{n} \frac{|\mathrm{t_i}|}{|\mathrm{t}|} \log_2 \left(\frac{|t_i|}{|t|} \right) \quad (4)$$

where $|t_i|$ is the number of observations that belong to class i in a specific leaf and $|t|$ gives the total number of instances in the leaf.

Another branching method, **Random forest (RF)**, allows a large number of classification trees to grow based on a CART tree using random bootstrap samples by replacement from the initial. In the training process, the algorithm produces enumerous decision trees using bootstrap. The decision mechanism of the RF works by voting since there are many trees in the forest, each one makes a decision. The average of these tree votes is used for classification. If we consider $f^b(x)$ as b^{th} decision tree, and B many trees in the forest or decision function becomes

$$\widehat{f}^b(x) = \frac{1}{B}\sum_{i=1}^{B} f^b(x) \quad .(5)$$

On the other hand, **XGBoost (XGB)**, a CART tree-based gradient-boosting approach, offers the advantages of speed, preventing overfitting, and handling missing observations, binary variables, and frequent zero values. It is another tree-based approach that considers multiple classification decisions. Unlike RF, all trees are generated sequentially. Each tree in XGB is adjusted by learning from the mistakes accrued in the previous tree. In each step, the algorithm constructs weak trees to avoid overfitting. The model is trained in an additive manner. Let $\widehat{y}_i^{(t)}$ be the predicted value of the ith sample at the tth iterations with

$$\widehat{y}_i^{(t)} = \sum_{k=1}^{K} f_k(x_i) = \widehat{y}_i^{(t-1)} + f_t(x_i) \quad (6)$$

where $\widehat{y}_i^0 = 0$, and $f_t(x_i) = \widehat{y}_i^{(t)} - \widehat{y}_i^{(t-1)}$. These adjustments are made by minimizing

$$L^t = \sum_{t=1}^{n} I\left(y_i, \widehat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad (7)$$

Here, Ω penalizes the complexity of each tree, and it is defined as $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda||w||^2$.

K-nearest neighbor (KNN) is a memory-based method that does not require any model fitting, and offers the advantage of no assumptions placed on the data apart from clustering with respect to distance metric. Given a testing point, x_{ij}, belonging to ith observation and jth feature in the sample space, we can find the closest k points, x_{ij}, $i = 1, 2, \ldots, n$, based on a distance function used. Then, we assign the instance (x_{ij}) to a suitable class by the use of majority vote over all the k neighbors. The closeness between two instance can be measured using the measure, $dist(x_1, x_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2}$.

A **multi-layer perceptron neural network (MLP)** is an extension of neural network, and is typically composed of an input layer, one or more hidden layers, and an output layer. The data samples initiate the system with the input layer that has a number of features times neurons. The first layer delivers the information to the next layer, the hidden layer. After that, each neuron receives the information from the previous layer. In each transaction, numerical information is multiplied by a weight, and all the connected neurons to the neuron in the next layer transfer their calculations which are aggregated along with a bias term. An activation function aids to transfer the calculations to the next layer till the aggregate value and the bias exceed a certain threshold. Similar estimations are employed in the output layer, and with a sigmoid activation function classification is occurred. If w_{jk}^l is the weight of the link from the kth neuron in the $(l-1)^{st}$ layer to the jth neuron in the lth layer and b_j^l is the bias term. Then, a_j^l , the activation of the jth neuron in the lth layer becomes

$$a_j^l = \sigma\left(\sum_k w_{jk}^l a_k^{l-1} + b_j^l\right) \quad (8)$$

The back-propagation algorithm is employed to adjust the weights by using a cost function in the training stage given as

$$C = \frac{1}{2n}\sum_x ||y(x) - a^L(x)||^2 \quad (9)$$

where L is the number of layers, $a^L(x)$ is the vector of activation function's output and $y(x)$ is the vector of target values for the n observation in the training set.

Due to the variety in the type of features in calculating traditional scores, we implement a **Generated Ensemble (GE)** algorithm to improve the accuracy power of the use of ML techniques. This approach produces the probabilities for each instances with respect to these eight ML algorithms. The average of these eight probabilities are classified with regard to a prespecified threshold value which converts the dependent variable to a binary one.

2.3. Performance measures and cross validation

Variation in the characteristics of the features of an ML method may cause over- or under-estimation. Accuracy must be achieved through data transformation techniques, such as Cox-Box and Min-Max, which enable algorithms to reduce interactions among variables and prevent domination of certain variables over others. Among eight selected ML algorithms, data transformation results in an influential improvement only on logistic regression, support vector machine and K-nearest neighbor techniques.

Prior to implementing the ML algorithms, we generate five random samples from both the original and transformed data sets by attaining the mortality percentage as observed in the binary dependent variable (ICU death rate- Exitus). This ensures that every possible outcome in the data set would be included, thus reducing limitations that can arise from small sample size. Then, for each random sample of five runs, with keeping the characteristics of the dependent variable to be the same, we create a training and a test groups having partitions of 80% and 20%, respectively. To elaborate more, we ensure the label balancing, by segregating the data with respect to its exitus states (1s or 0s). Afterwards, 80%–20% partition of 1s are separated as train and test (train_1, test_1) attaing the same to be done for the set of 0s (train_0, test_0). Then, train_0 and train_1 sets are merged to create balanced one train dataset as well as the same is done for the test parts. In the ultimate test data set ends up with 41 observations, assuring a balanced structure in dependent variable. Traditional methods use the complete data set and their results constitute the base for comparison. It should be noted that separating the data set as train and test parts reduces the number of observations. However, the randomly selected sets of train and test groups processed in the ML algorithms, assures the inclusion of all possible selection as well. As next, the parameters required in the ML algorithms are optimized using Grid Search method. A ten-fold cross validation is implemented to the train set during the process of grid search parameter optimization. The degree of consistency between the observed and predicted values (i.e., true positives and true negatives) determines the measure of accuracy, and the aim with each ML method is to avoid overestimation (classified as Class 0 with the "zero" numeral) or underestimation (classified as Class 1 with the "one" numeral) of the binary dependent variable [26–30,32].

The AUC-ROC curve generated for each ML method and ensemble of those shows the trade-off between the true positive and true negatives rates; that is, the incorrectly classified "ones" in Class 1 to the total number of cases in that class (i.e., Type I error or Specificity) and the same proportion for falsely detected "zeros" (i.e., Type II error or Sensitivity). We use these findings to determine the accuracy of our ML methods [33]. It is important to reiterate that, we apply each ML method using the same variables that are used in calculating APACHE II, SAPS II and SOFA scores for ICU patients. For each ML method we analyze, the means for AUC, Type I error rate, and Type II error rate are calculated

across the five independent runs. The ML method that yields the best performance initially is then tested further, with focus on the features that are considered significant in predicting sepsis mortality. The algorithms conducted on training data set enables us to determine the optimized parameters whose values are used in test data sets. Over five independent trials, the average values of performance indicators, *i.e.*, AUC, specificity, and sensitivity rates yield the final output of the analyses. Additionally, the features having a direct impact on the prediction can also be listed based on the best performing ML method. We illustrate in Fig. 1 aforementioned steps of implementing selected ML methods.

2.4. Statistical analysis

The analysis to understand the behavior of the variables, we calculate descriptive statistics (*e.g.*, mean, standard deviation, minimum and maximum values, coefficient of variation). Binary variables are represented in terms of percentages with their standard deviations in paranthesis. The data set collected in our study do not have any missing values, as the recordings are made with respect to the ingredients required to calculate APACHE II, SAPS II and SOFA. The data distribution in classes are unbalanced. The proportion of sepsis mortality in overall set is taken as threshold to create the random samples from the original data when applying the ML methods. To investigate whether any triggering disease has higher influence on mortality, Fisher Exact test is employed whose outcome indicates that Chronic renal failure and lung cancer have significant impact on ICU death rate with p-values 0.045 and 0.001, respectively. The Chi-square goodness of fit test and Jargue-Bera (JB) tests are employed to verify the Normal distribution. The JB test illustrates that the variables Temperature, Urine Output, Bicarbonate, Sodium follow Normal distribution whose p-values are marked in boldface in Table 1. All quantitative variables are normalized using Box-Cox transformation. The association between features are analyzed using Pearson correlation and point bi-serial correlation. The data set is re-scaled using min-max transformation to eliminate the discrepancy in units. This is also an important factor on bringing the comparisons on a unique measure. AUC-ROC scores are calculated based

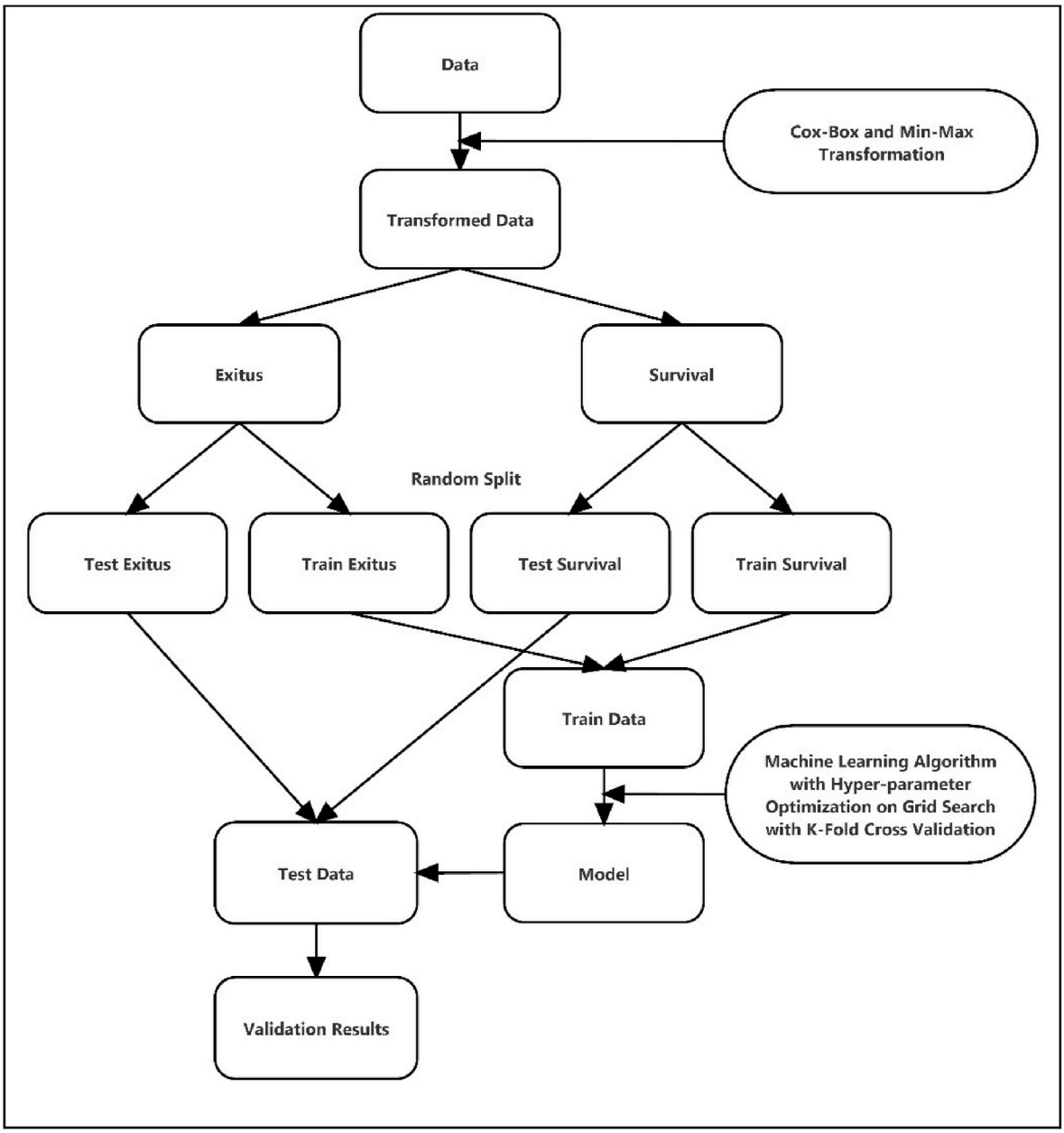

Fig. 1. The implementation of the ML algorithms.

on the predictions obtained using each ML methods and traditional scores whose comparisons are done also using AUC-ROC curves.

3. Results

Table 1 outlines the descriptive statistics for the variables employed in the study. The common variables required to calculate APACHE II, SAPS II and SOFA are listed in the second column of the table. The mean with standard deviations in the paranthesis, minimum, maximum, and the coefficient of variation (CoV) of quantitative features, as well as the percentages for binary variable (1 when death occurs), give first insight on the variables. The high average age of 75 is due to an extreme value (100). The coefficient of variation (CoV) illustrating the homogeneity of the observations is the highest in Creatinine (68%). Even though the rates of Organ Failure (40.50%) and Chronic Disease (49%) are remarkable, slightly less than half of the cases end up with ICU sepsis death rate (33.50%). Additionally, no significantly triggering disease is found. The p-value$<$0.05 is accepted to be statistically significant for all tests. A *t*-test states no evidence to claim that the mortality rate changes with gender ($p < 0.001$).

The distribution of the observations for each variable is presented in Fig. A1. The age distribution of patients is mostly densed between 60 and 90C, whereas the heart rates show more tendency into the direction of low values. Systolic BP is highly concentrated around 78 and 97 (mmHg). BUN, Potassium, Bilirubin, Respiration rate are mostly left skewed having the aggregation in the first categories. On the other hand, Temperature, Urine Output, Sodium, Bicabonate, WBC and MAP show more compiled structure around their mean values. PaO_2 has long tailed behavior, whereas, Glaskow Coma score and FiO_2 show more uniformity in distributions (Fig. A1). The significant dependence between features is outlined as follows: Creatine and Acute renal failure (%69), Glasgow coma score and ICU death rate (%52). On the other hand, FiO_2 and PaO_2, Arterial pH and Glasgow coma score, BUN and Creatinine, Arterial pH and ICU death rate depict correlations around 30%, and the rest of the pairwise dependence is below 20%.

The AUC-ROC curves for each score and algorithms under every resampling and transformation processes are calculated and plotted to expose the balance between true positive (TP) and false positive (FP) rates. Table 2 illustrates accuracy indicators, AUC, specificity, and sensitivity, to analyze with respect to ML algorithms as well as traditional scores. In case of considering solely APACHE II, SAPS II, and SOFA scores as mortality measure, we see that these yield AUC ratios of 82.60%, 78.81%, and 80.90%, respectively. Taking into account using the same features, ML methods are found to give very promising predictive power for sepsis mortality. Among eight algorithms, MLP gives the best performance (AUC value is 84.45%) compared to the ones obtained from APACHE II and SAPS II. MLP which stems from NN show better accuracy in estimating APACHE II and SAPS II, because of the variation in hidden layers having influence on the parameter estimations. On the other hand, traditional SOFA score whose AUC is 9% higher than the ones obtained by RF yielding the best performance (AUC of 73.47%) among other algorithms to estimate the mortality. It should be noted that even the worst performing ML methods (such as KKN in APACHE II, GNB in SAPS II, and SOFA) remain or are close to and consistent with the literature which accounts AUC$>$70% to be an acceptable level for a good prediction level [34–41]. Fig. 2 illustrates AUC values for traditional scores (left panel) against the AUCs of the best performing ML methods (middle panel) and the ensemble of ML algorithms (left panel). Comparing the accuracy of ML algorithms, we find MLP is the best choice in quantifying the sepsis mortality risk whose AUC is shown in Fig. 2d and 3 e. It gives much higher TP rates against sensitivity indicator compared to AUCs of APACHE II and SAPS II (Fig. 2a and 3 b). On the other hand, the AUC of RF (Fig. 2f) compiled with respect to SOFA inputs competes with the AUC graph of the score itself (Fig. 2c). Meanwhile, the specificity and sensitivity ratios (Table 2) agree with the AUC performances. Application of the GE algorithm show the best development in APACHE II score (Fig. 2g) and the AUC curve is recognizable improved in SOFA (Fig. 2i). The performance results on the GE application shown in Table 3 prove that ensembled model functions the best in APACHE II predictors compared with the other implemented ML algorithms and traditional scoring methods. Also, its specificity reaches to its maximum value, 1.0. The sensitivities of SAPS II and SOFA in ensembled algorithm obtained with their related features are improved compared to the best fitting ML algorithms. It is worthwhile to point out that the standard deviations in GE algorithm are reduced dramatically.

Another aspect in employing ML in comparison to the conventional scores is its reliability due to the validation process at which the sample is randomly assigned to each class to reduce the bias in prediction. In case of large sample size, influential factors or variables which improve the performance scores may change. Though small set of observations fail to distinguish the mortality rate reliably, a large sample size may not always be available or may require a long time period for realizing the data set. On the other hand, in line with some literature on the

Table 2
Prediction power of ML methods and classical scores with their standard deviations (in paranthesis).

	Performance indicators (standard deviations) using ML algorithm								
	APACHE II			SAPS II			SOFA		
ML	AUC	Specificity	Sensitivity	AUC	Specificity	Sensitivity	AUC	Specificity	Sensitivity
MLP	**0.8445** (0.0976)	0.8888 (0.0908)	0.7714 (0.1864)	**0.8524** (0.0631)	0.9333 (0.1326)	0.7714 (0.212)	0.7188 (0.1646)	0.8518 (0.1386)	0.5857 (0.3698)
SVM	0.8127 (0.126)	0.9111 (0.0846)	0.7142 (0.226)	0.7635 (0.1287)	0.9555 (0.1606)	0.5714 (0.226)	0.7339 (0.1812)	0.8962 (0.062)	0.5714 (0.3644)
LR	0.781 (0.1003)	0.9333 (0.0966)	0.6285 (0.1566)	0.8236 (0.0786)	0.9185 (0.177)	0.7285 (0.1864)	0.7236 (0.208)	0.9185 (0.0332)	0.5285 (0.399)
RF	0.8387 (0.117)	0.9629 (0.0908)	0.7142 (0.175)	0.7707 (0.106)	0.9555 (0.062)	0.5857 (0.1864)	**0.7347** (0.194)	0.9407 (0.0994)	0.5285 (0.3442)
XGB	0.7985 (0.1312)	0.9111 (0.1444)	0.6857 (0.2168)	0.7588 (0.0547)	0.8888 (0.1386)	0.6285 (0.1566)	0.6895 (0.1448)	0.8074 (0.0966)	0.5714 (0.3644)
KNN	0.7667 (0.0758)	0.9333 (0.062)	0.6000 (0.163)	0.6172 (0.0961)	0.9629 (0.0908)	0.2714 (0.2348)	0.7165 (0.1837)	0.9185 (0.169)	0.5142 (0.3558)
DT	0.7553 (0.1187)	0.8962 (0.1218)	0.6142 (0.329)	0.7514 (0.0866)	0.874 0.1444)	0.6285 (0.3098)	0.686 (0.2336)	0.8148 (0.1172)	0.5571 (0.4092)
GNB	0.7805 (0.0938)	0.9037 (0.0846)	0.6571 (0.1196)	0.6495 (0.0825)	0.9703 (0.062)	0.3285 (0.163)	0.6823 (0.1722)	0.8074 (0.1426)	0.5571 (0.3834)
	Performance indicators (standard deviations) using traditional scores								
	APACHE II			**SAPS II**			**SOFA**		
AUC	0.826 (0.0554)			0.7881 (0.0906)			0.809 (0.0459)		

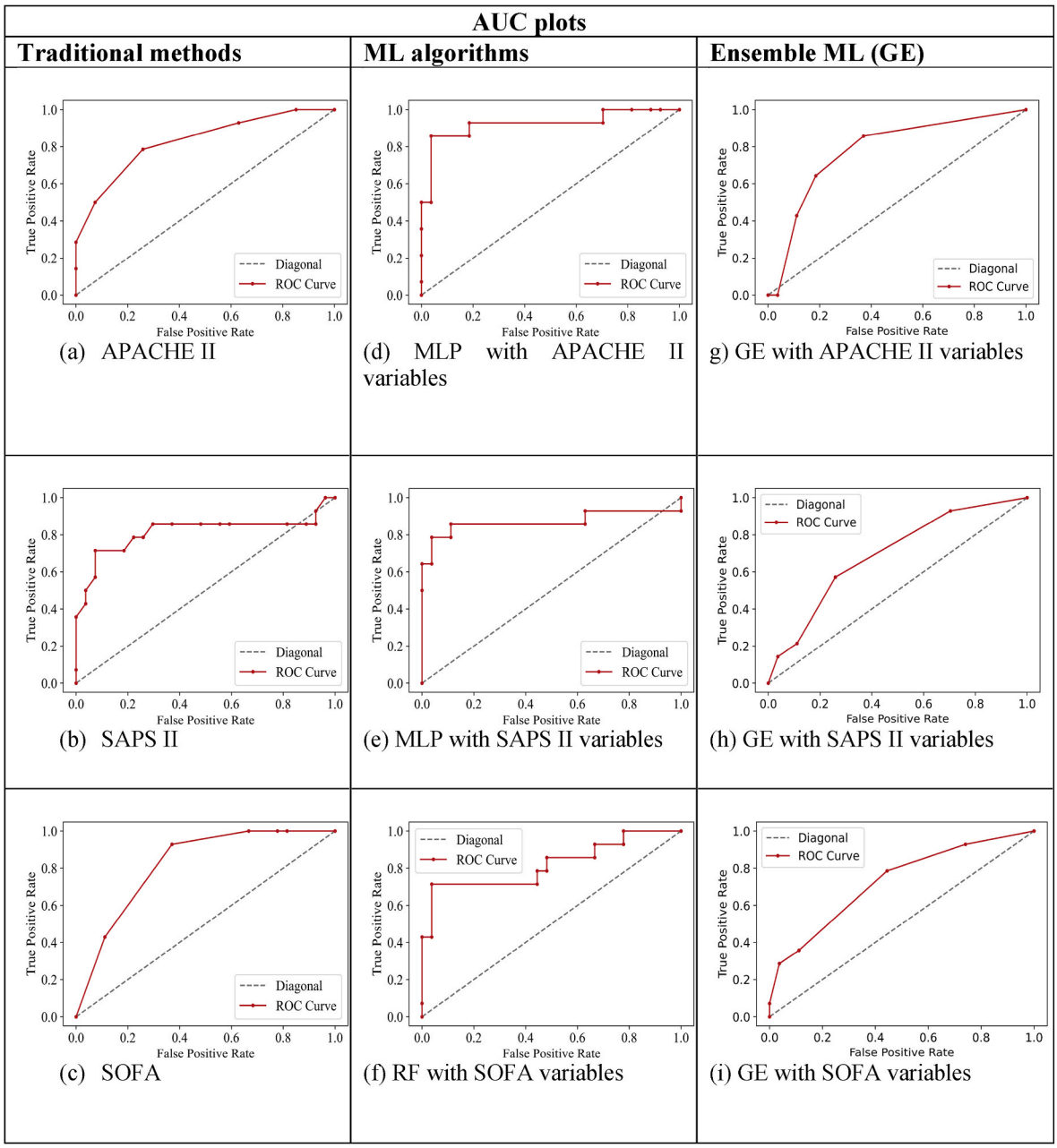

Fig. 2. AUCs of traditional scores (left panel), the best performing ML algorithms (middle panel) and ensemble model (right panel).

Table 3
Prediction power of generated ensemble (GE) algorithm and classical scores (standard deviation).

	APACHE II	SAPS II	SOFA
AUC	**0.8730** (6.6628. E−4)	0.8413 (1.7161. E−3)	0.7307 (6.5575. E−3)
Sensitivity	0.6717 (4.3365. E−5)	**0.7977** (1.2524. E−3)	**0.7766** (3.0534. E−3)
Specificity	**1.0000** (0.0000)	0.8614 (5.8747. E−3)	0.6115 (1.0384. E−2)
AUC (scoring methods)	0.8260 (5.5400. E−2)	0.7881 (9.0600. E−2)	0.8090 (4.5900. E−2)

implementation of ML for small sample cases with satisfactory prediction powers [42,43], our results are found to be convincing on the reliability of implemented ML methods. On the other side, we show that, the consequentially data transformations (Box-Cox and Min-Max) eliminates the influence of extreme values, prevent some features outweighing the others completely with a smaller range of measurement, and assures normality assumption in order to contribute to the efficiency of the ML systems.

Besides accuracy performance, parameter based outlook for each algorithm explains the convergence ability of the selected methods as presented in Fig. 3. Except RF, all algorithms yield better performance when APACHE II features are used as predictors. The regularization parameter λ for LASSO on LR, has an increasing effect on the performance of the model (Fig. 3a). and yields the highest AUC when λ=*2*. In MLP, 13 different layers and neuron combinations are examined (Fig. 3b). When the number of layers is increased, the complexity of the model also increases. The model with two layers and 100 neurons gives the most promising result (AUC of 82.58%). RF gives more robust results for SOFA features. The model performance is improved as the number of trees are increased and volatility is decreased (Fig. 3c). The KNN has optimal 11 neighbors for the best AUC (Fig. 3d). The results observed in

Fig. 3. Parameterization and specifics required in the computation of ML algorithms.

Fig. 3e depict that XGB does not require a sophisticated model and yields the highest performance with five trees (AUC of 84.30%). In DT algorithm, maximum depth is decided to be the base parameter which limits the branching mechanism. As the branching increases, the model fits the training data more, and creating an overfitting problem. When the parameter is fixed to 2, the model performs better with AUC of 84.93% (Fig. 3f). GNB calculates the probabilities on the training dataset, and there is a chance that a categorical feature of an instance in the test dataset can be different from the training dataset, and it causes the probability to be zero. To prevent having zero probability, the variance smoothing parameter is chosen to range between [1.e-14,1.e-1]. Fig. 3g shows that small values affect the model positively and the performance does not change after the parameter value of 1.e-06 whose AUC is 84.39%. In SVM, as the regularization is increased, the rate of tolerance for misclassification decreases. It also controls the hyperplane to be a non-linear separation border. Fig. 3h depicts that the optimal value for the parameter becomes 2 (AUC of 77.68%).

4. Discussion

The investigation on the prediction performance of multiple ML methods indicates that MLP handles the influence of variables on sepsis mortality better than other ML methods, as well as APACHE II and SAPS II. MLP which stems from NN show better accuracy in estimating APACHE II and SAPS II, because they create sophisticated and complex structures by creating the hidden layers which captures all invisible connections. In the literature many of the ML studies [20,22,23,26,43], have agreement on the efficiency of NN originated algorithms. Traditional ICU scores, which are easy and brief in the application, do not show a larger discrepancy in mortality prediction power compared to MLP. Nemati et al. [1] also state that an ML algorithm can precisely predict the onset sepsis of the patients within 4–12 h before the clinical diagnosis. Zhang et al. [17]. propose a severity score system based on the data retrieved from the Medical Information Mart for Intensive Care III (MIMIC III) to show the efficiency of employing ML on detecting sepsis-3 patients in ICU. Similarly, Kong et al. [18] justifies our findings on the

MIMIC III data set to detecting sepsis-3 mortality in the utilization of ML methods in predicting mortality even though the defining variables are not restricted to the traditional score inputs as we propose in this paper. Another specialty in this study is on data transformation (Box-Cox and Min-Max) and parameter optimization (Grid Search), which are shown to improve the accuracy on the estimation of sepsis mortality. As an important output of MLP algorithm, the most influential features on the mortality are found to be Heart rate, Glasgow Coma Score, Urine output, Creatinine, and the existence of comorbidities. This indicates that ML offers an important advantage to distinguish the variables which are more critical on mortality occurrence. The algorithms applied to the data whose variables are the main ingredients of calculating the traditional scores yield the estimates for sepsis mortality as an outcome. Based on the performance measures, we could make this inference on the success of MLP. Such specific information cannot be retrieved with traditional scores. Hence, we show how ML can be used as a clinical contribution for predicting mortality.

Studies show that as conventional scores APACHE II and SAPS II yield similar power in prediction, even though the calibration of these is not straightforward, which brings experts to take into account multiple scorings in their decision making as imposed in Refs. [10,44]. To tackle the efficiency problems along with handling the continuously flowing data, which requires adequate reaction at a short time span together with reducing the mortality or the conditions contributing to mortality, our findings for implementing ML methods in sepsis cases is supported by the literature [10,45–47]. The neural network, random forest, support-vector-machine and tree-based methods are also found to perform as good as APACHE II and III scores in estimating the mortality and they require fewer variables to quantify these risks [25–27,48] as well as SOFA score [27–29,48]. Contrary to APACHE III and SAPS II, SOFA concentrates on the failure of the organs, that is, the main contributing factors are the organ functions and their indicators whose values are entered to the web-based software yielding the mortality risk. Even, ML methods did not over perform traditional SOFA score as it is also emphasized by Zheng et al. [22] and Zhao et al. [49], their performance is in accordance with good performance range in predicting the mortality. These findings justify our motivation on implementing ML techniques to determine intensive care indicators for decision making.

The main limitation in this study is mainly the small sample size compared to many other studies. Even though the larger data is known to provide more convincing and reliable results, there is no fixed minimum sample size stated in literature for applying ML methods. The sample size of 200 in this study may be taken not sufficiently large in consideration of the diversity in comorbidities of the patients. Nevertheless, the convincing results on the ML power indicators, and literature with small sample applications [16–18,32] relieve this concern. Another drawback is that the observations represent the first day at ICU and do not reflect the progression of patients on the treatments and/or mortality. As the gradual changes in the data by time are not recorded, this study lacks of such comparative analyses.

5. Conclusion

Sepsis is one of the leading mortality cause, especially in ICU patients whose state of health requires monitoring the continual flow of the observations from multiple sources and require quick and instant actions by experts. Additional to the studies questioning if ML contributes to the prediction of ICU mortality compared to traditional scoring systems, we aim to investigate whether prediction of sepsis mortality is more efficient and faster with automatized methods (such as ML algorithm) and if these can be used as a supporting mechanism for classical scoring methods. Besides, local and regional factors may also have influence on the mortality risk which can be easily implemented and processed by ML algorithms which will also enable users to develop a custom-tailored interface to evaluate and predict mortality risk. In traditional scores, all variables contribute to the calculation of the score, whereas, ML methods have ability to distinguish which of the variables have more influence on the sepsis mortality. The feature selection property of ML allows us to choose the most effective ones. The intermediate information is the most decisive in clinical treatments and has an important impact on the prognosis in ICU. This retrospective observational study compares eight ML methods with traditional ICU scores (APACHE II, SAPS II, and SOFA) for predicting sepsis mortality which is an essential input for risk adjustment and decision making at the ICU. The main contribution of this study is to depict the inevitable implementation of artificial intelligence in ICU decision mechanisms to show how efficient and practical to utilize ML methods to predict the mortality of sepsis patients.

Funding

There is no funding sources provided from any sources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Authors are grateful to the Intensive Care Unit at Acibadem Kadiköy Dr. Sinasi Can Hospital and Ethics Committee at Acibadem University.

Appendix. Supplementary materials on distributions and hyperparametrization

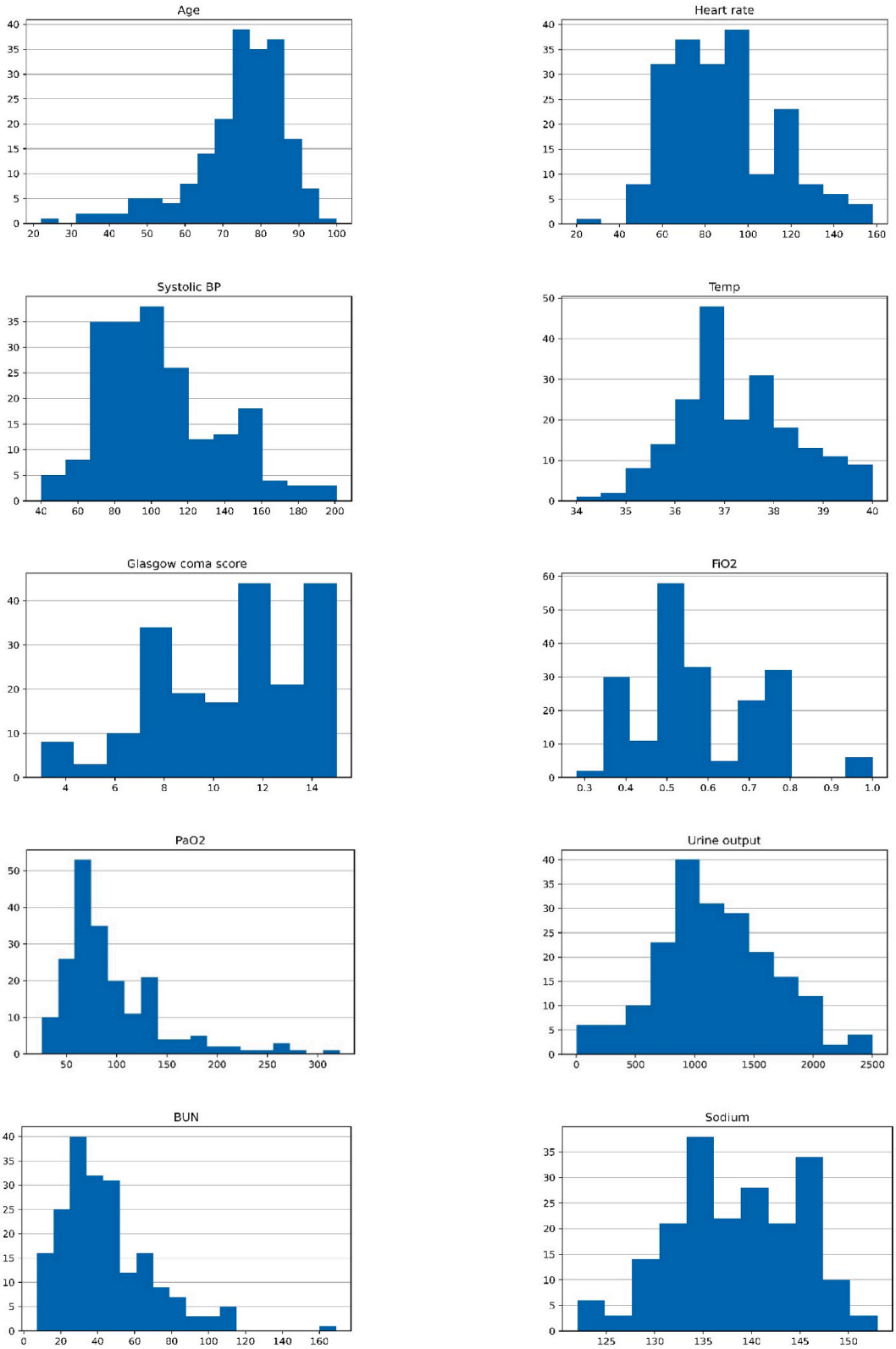

Fig. A1. The distributions of the features used in algorithms

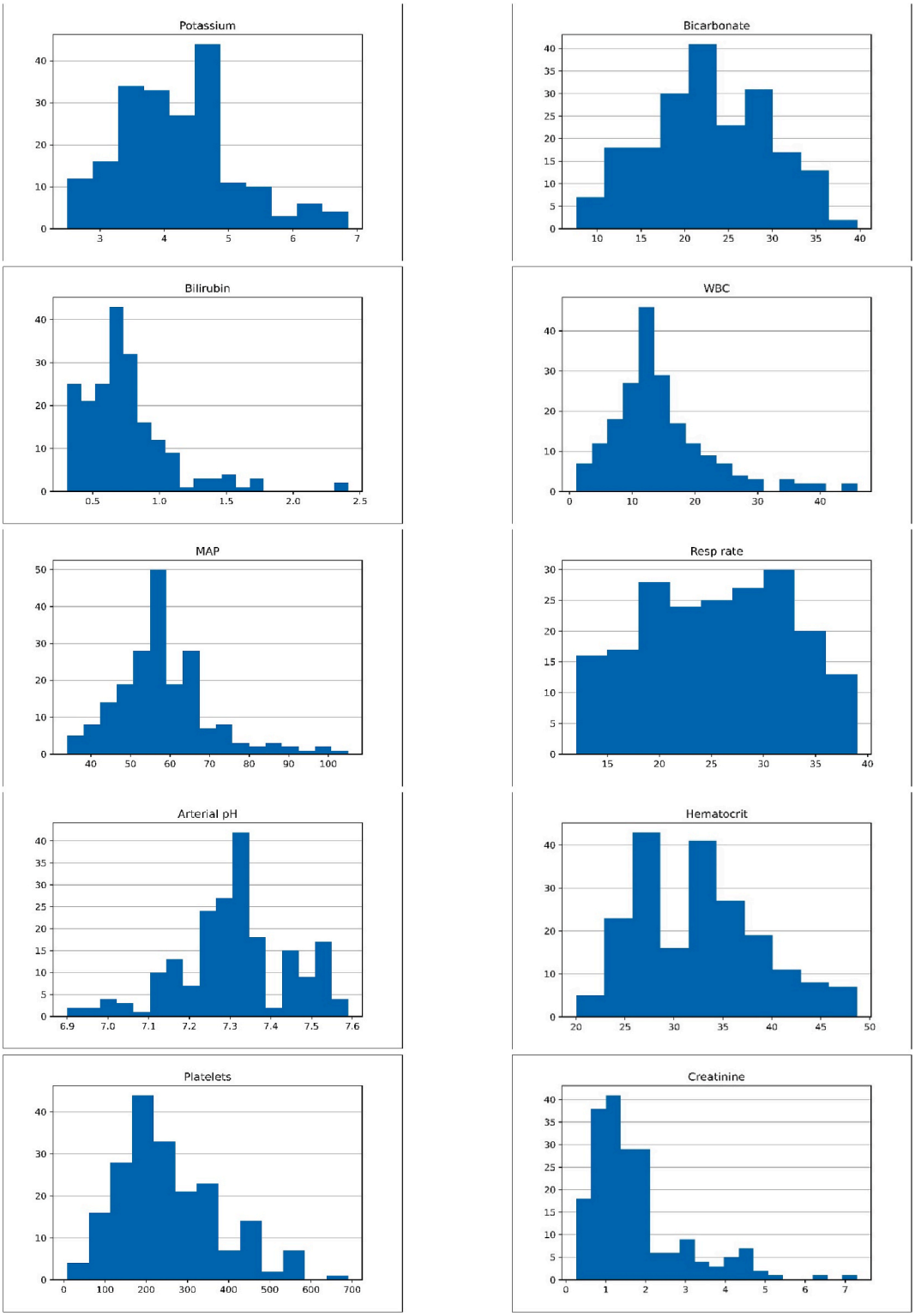

Fig. A1. (*continued*).

Table A1
Algorithms and their specifications by Grid Search parametrizations

ML	Specifications	Parameters (by Grid Search)
RF	Number of trees:	20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120
	Bootstrap:	"True", "False"
	Criterion:	"Gini", "Entropy"

(*continued on next page*)

Table A1 (*continued*)

ML	Specifications	Parameters (by Grid Search)
	Minimum impurity split:	1.00E-07
	Minimum Samples in a leaf:	1, 2, 3
	Maximum depth of a tree:	2, 3, 5, 7, 10, None
	Maximum features:	"Auto", "Sqrt"
DT	Criterion:	"Gini", "Entropy"
	Minimum Samples in a leaf:	1, 2,3, 4, 10, 20, 50, 100
	Maximum depth of a tree:	1, 2, 3, 4, 5, 8, 10, 11, 12, 13
	Maximum features:	"Auto", "Sqrt", "log2″
	Splitter:	"Best", "Random"
	Complexity parameter:	0.0, 0.1, 0.2
XGB	Number of trees:	20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120
	Sampling:	"Gradient based", "Uniform"
	Maximum depth:	2, 3, 5, 7, 10
	Tree method:	"Auto", "Exact"
	Minimum child weight:	2, 5
	Gamma:	0.06, 0.01, 0.1, 0.2, 0.5
KNN	Algorithm:	"Auto"
	Weights:	"Uniform", "Distance"
	Number of neighbors:	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
	Power of Minkowski metric:	1, 2, 3, 4
MLP	Number of layer and neuron:	(20),(30),(50),(100),(20,20),(30,30),(50,50),(100,100), (20,20,20),(30,30,30),(50,50,50),(100,100,100)
	Solver:	"Lbfgs"
	Activation:	"Logistic", "Relu", "Tanh", "Identitiy"
	L2 penalty parameter:	0.01, 0.001, 0.0001
	Tolerance for stopping criterion:	0.01, 0.001, 0.0001, 0.00001
	Validation fraction:	0.1, 0.2
	Maximum iteration:	100, 500, 1000, 2500, 5000
SVM	Regularization parameter:	0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 4
	Tolerance for stopping criterion:	1E-2, 1E-3, 1E-4, 1E-5, 1E-6, 1E-8
	Kernel:	"Rbf", "Poly"
	Maximum iteration:	100, 250, 500, 1000, 1500, −1
	Decision function:	"ovo", "ovr"
LR	Regularization parameter:	0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 4
	Tolerance for stopping criterion:	1E-3, 1E-4, 1E-5, 1E-6
	Penalty:	l2, None
	Solver:	"Lbfgs"
	Maximum iteration:	100, 250, 500, 1000, 1500
GNB	Variance smoothing:	1E-2,1E-3,1E-4,1E-5,1E-6,1E-7,1E-8,1E-9,1E-10,1E-11, 1E-12

References

[1] Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit Care Med 2018. https://doi.org/10.1097/CCM.0000000000002936.

[2] Knaus WA, Draper EA, Wagner DP, Zimmerman JE. Apache II: a severity of disease classification system. Crit Care Med 1985. https://doi.org/10.1097/00003246-198510000-00009.

[3] Le Gall JR. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA, J Am Med Assoc 1993. https://doi.org/10.1001/jama.270.24.2957.

[4] Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive Care Med 1996. https://doi.org/10.1007/BF01709751.

[5] Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in the ICU: can we do better? Results from the Super ICU Learner Algorithm (SICULA) project, a population- based study. Lancet Respir Med 2015.

[6] Gambus P, Shafer SL. Artificial intelligence for everyone. Anesthesiology 2018. https://doi.org/10.1097/ALN.0000000000001984.

[7] Lin CS, Chang CC, Chiu JS, Lee YW, Lin JA, Mok MS, et al. Application of an artificial neural network to predict postinduction hypotension during general anesthesia. Med Decis Making 2011. https://doi.org/10.1177/0272989X10379648.

[8] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019. https://doi.org/10.1056/NEJMra1814259.

[9] Alexander JC, Joshi GP. Anesthesiology, automation, and artificial intelligence. Baylor Univ Med Cent Proc; 2018. https://doi.org/10.1080/08998280.2017.1391036.

[10] Côté CD, Kim PJ. Artificial intelligence in anesthesiology: moving into the future. Univ Toronto Med J; 2019.

[11] Connor CW. Artificial intelligence and machine learning in anesthesiology. Anesthesiology 2019. https://doi.org/10.1097/ALN.0000000000002694.

[12] Mathis MR, Kheterpal S, Najarian K. Artificial intelligence for anesthesia: what the practicing clinician needs to know more than black magic for the art of the dark. Anesthesiology 2018. https://doi.org/10.1097/ALN.0000000000002384.

[13] Desautels T, Das R, Calvert J, Trivedi M, Summers C, Wales DJ, et al. Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: a cross-sectional machine learning approach. BMJ Open 2017. https://doi.org/10.1136/bmjopen-2017-017199.

[14] Tighe PJ, Harle CA, Hurley RW, Aytug H, Boezaart AP, Fillingim RB. Teaching a machine to feel postoperative pain: combining high-dimensional clinical data with machine learning algorithms to forecast Acute postoperative pain. Pain Med (United States); 2015. https://doi.org/10.1111/pme.12713.

[15] Thottakkara P, Ozrazgat-Baslanti T, Hupf BB, Rashidi P, Pardalos P, Momcilovic P, et al. Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. PLoS One 2016. https://doi.org/10.1371/journal.pone.0155705.

[16] undefined Johnson A, Pollard T, Shen L, Li-Wei H, data MF-S. MIMIC-III, a freely accessible critical care database. NatureCom n.d; 2016.

[17] Zhang K, Zhang S, Cui W, Hong Y, Zhang G, Zhang Z. Development and validation of a sepsis mortality risk score for sepsis-3 patients in intensive care unit. Front Med 2021;7:1142.

[18] Kong G, Lin K, Hu Y. Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. BMC Med Inf Decis Making 2020;20. https://doi.org/10.1186/S12911-020-01271-2.

[19] Anitescu C, Atroshchenko E, Alajlan N, Rabczuk T. Artificial neural network methods for the solution of second order boundary value problems. Comput Mater Continua (CMC) 2019;59:345–59. https://doi.org/10.32604/CMC.2019.06641.

[20] Wernly B, Mamandipoor B, Baldia P, Jung C, Osmani V. Machine learning predicts mortality in septic patients using only routinely available ABG variables: a multi-centre evaluation. Int J Med Inf 2021;145:104312. https://doi.org/10.1016/J.IJMEDINF.2020.104312.

[21] Yao RQ, Jin X, Wang GW, Yu Y, Wu GS, Zhu YB, et al. A machine learning-based prediction of hospital mortality in patients with postoperative sepsis. Front Med 2020;7:445.

[22] Zeng Z, Yao S, Zheng J, Gong X. Development and validation of a novel blending machine learning model for hospital mortality prediction in ICU patients with Sepsis. BioData Min 2021;14:1–15.

[23] Zhang Z, Pan Q, Ge H, Xing L, Hong Y, Chen P. Deep learning-based clustering robustly identified two classes of sepsis with both prognostic and predictive values. EBioMedicine 2020;62:103081. https://doi.org/10.1016/J.EBIOM.2020.103081.

[24] Samaniego E, Anitescu C, Goswami S, Nguyen-Thanh VM, Guo H, Hamdia K, et al. An energy approach to the solution of partial differential equations in computational mechanics via machine learning: concepts, implementation and applications. Comput Methods Appl Mech Eng 2020;362:112790. https://doi.org/10.1016/J.CMA.2019.112790.
[25] Kim S, Kim W, Woong Park R. A comparison of intensive care unit mortality prediction models through the use of data mining techniques. Healthc Inform Res 2011. https://doi.org/10.4258/hir.2011.17.4.232.
[26] Meiring C, Dixit A, Harris S, MacCallum NS, Brealey DA, Watkinson PJ, et al. Optimal intensive care outcome prediction over time using machine learning. PLoS One 2018. https://doi.org/10.1371/journal.pone.0206862.
[27] Aperstein Y, Cohen L, Bendavid I, Cohen J, Grozovsky E, Rotem T, et al. Improved ICU mortality prediction based on SOFA scores and gastrointestinal parameters. PLoS One 2019. https://doi.org/10.1371/journal.pone.0222599.
[28] Jeleazcov C, Egner S, Bremer F, Schwilden H. Automated EEG preprocessing during anaesthesia: new aspects using artificial neural networks. Biomed Tech 2004. https://doi.org/10.1515/BMT.2004.025.
[29] Peker M, Şen B, Gürüler H. Rapid automated classification of anesthetic depth levels using GPU based parallelization of neural networks. J Med Syst 2015. https://doi.org/10.1007/s10916-015-0197-3.
[30] Singer M, Deutschman CS, Seymour C, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). JAMA, J Am Med Assoc 2016. https://doi.org/10.1001/jama.2016.0287.
[31] APACHE II Score - MDCalc n.d. https://www.mdcalc.com/apache-ii-score (accessed December 29, 2021).
[32] Aminiahidashti H, Bozorgi F, Montazer SH, Baboli M, Firouzian A. Comparison of Apache II and SAPS II scoring systems in prediction of critically ill patient's outcome. Arch Acad Emerg Med 2019. https://doi.org/10.22037/emergency.v5i1.11282.
[33] Han J, Kamber M, Pei J. Data mining: concepts and techniques. 2012. https://doi.org/10.1016/C2009-0-61819-5.
[34] Sungurtekin H, Gürses E, Balci C. Evaluation of several clinical scoring tools in organophosphate poisoned patients. Clin Toxicol 2006. https://doi.org/10.1080/15563650500514350.
[35] Khwannimit B, Geater A. A comparison of Apache II and SAPS II scoring systems in predicting hospital mortality in Thai adult intensive care units. J Med Assoc Thail 2007.
[36] Jones AE, Trzeciak S, Kline JA. The Sequential Organ Failure Assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation. Crit Care Med 2009. https://doi.org/10.1097/CCM.0b013e31819def97.
[37] Alizadeh AM, Hassanian-Moghaddam H, Shadnia S, Zamani N, Mehrpour O. Simplified acute physiology score II/Acute physiology and chronic health evaluation ii and prediction of the mortality and later development of complications in poisoned patients admitted to intensive care unit. Basic Clin Pharmacol Toxicol 2014. https://doi.org/10.1111/bcpt.12210.
[38] Haddadi A, Ledmani M, Gainier M, Hubert H, De Micheaux PL. Comparing the Apache II, SOFA, LOD, and SAPS II scores in patients who have developed a nosocomial infection. Bangladesh Crit Care J 2014. https://doi.org/10.3329/bccj.v2i1.19949.
[39] Fernando SM, Tran A, Taljaard M, Cheng W, Rochwerg B, Seely AJE, et al. Prognostic accuracy of the quick sequential organ failure assessment for mortality in patients with suspected infection, A systematic review and meta-Analysis. Ann Intern Med 2018. https://doi.org/10.7326/M17-2820.
[40] Lemeshow S, Le Gall JR. Modeling the severity of illness of ICU patients: a systems update. JAMA, J Am Med Assoc 1994. https://doi.org/10.1001/jama.1994.03520130087038.
[41] Beck DH, Smith GB, Pappachan JV, Millar B. External validation of the SAPS II, Apache II and Apache III prognostic models in South England: a multicentre study. Intensive Care Med 2003. https://doi.org/10.1007/s00134-002-1607-9.
[42] P D, R C, L N, R G, F S, B G. 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 2012.
[43] Lukaszewski RA, Yates AM, Jackson MC, Swingler K, Scherer JM, Simpson AJ, et al. Presymptomatic prediction of sepsis in intensive care unit patients. Clin Vaccine Immunol 2008. https://doi.org/10.1128/CVI.00486-07.
[44] Tang CQ, Li JQ, Xu DY, Liu XB, Hou WJ, Lyu KY, et al. Comparison of machine learning method and logistic regression model in prediction of acute kidney injury in severely burned patients. Zhonghua Shaoshang Zazhi 2018. https://doi.org/10.3760/cma.j.issn.1009-2587.2018.06.006.
[45] Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. JMIR Med Informatics 2016. https://doi.org/10.2196/medinform.5909.
[46] Poncet A, Perneger TV, Merlani P, Capuzzo M, Combescure C. Determinants of the calibration of SAPS II and SAPS 3 mortality scores in intensive care: a European multicenter study. Crit Care 2017. https://doi.org/10.1186/s13054-017-1673-6.
[47] Ranta SOV, Hynynen M, Räsänen J. Application of artificial neural networks as an indicator of awareness with recall during general anaesthesia. J Clin Monit Comput 2002. https://doi.org/10.1023/A:1015426015547.
[48] Kang MW, Kim J, Kim DK, Oh KH, Joo KW, Kim YS, et al. Machine learning algorithm to predict mortality in patients undergoing continuous renal replacement therapy. Crit Care 2020;24:1–9.
[49] Zhao X, Shen W, Wang G. Early prediction of sepsis based on machine learning algorithm. Comput Intell Neurosci 2021:2021. https://doi.org/10.1155/2021/6522633.