Using machine learning in prediction of ICU admission, mortality, and length of stay in the early stage of admission of COVID-19 patients

Saadatmand, S.; Salimifard, K.; Mohammadi, R.; Kuiper, A.; Marzban, M.; Farhadi, A.

Using machine learning in prediction of ICU admission, mortality, and length of stay in the early stage of admission of COVID-19 patients

Sara Saadatmand[1] · Khodakaram Salimifard[1] · Reza Mohammadi[2] ·
Alex Kuiper[2] · Maryam Marzban[3] · Akram Farhadi[4]

Abstract

The recent COVID-19 pandemic has affected health systems across the world. Especially, Intensive Care Units (ICUs) have played a pivotal role in the treatment of critically-ill patients. At the same time however, the increasing number of admissions due to the vast prevalence of the virus have caused several problems for ICU wards such as overburdening of staff and shortages of medical resources. These issues might have affected the quality of healthcare services provided directly impacting a patient's survival. The objective of this research is to leverage Machine Learning (ML) on hospital data in order to support hospital managers and practitioners with the treatment of COVID-19 patients. This is accomplished by providing more detailed inference about a patient's likelihood of ICU admission, mortality and in case of hospitalization the length of stay (LOS). In this pursuit, the outcome variables are in

✉ Khodakaram Salimifard
salimifard@pgu.ac.ir

Sara Saadatmand
sara.saadatmand@mehr.pgu.ac.ir

Reza Mohammadi
a.mohammadi@uva.nl

Alex Kuiper
a.kuiper@uva.nl

Maryam Marzban
marzbanh@gmail.com

Akram Farhadi
ak.farhadi@gmail.com

[1] Computational Intelligence and Intelligent Optimization Research Group, Persian Gulf University, Bushehr 75169, Iran

[2] Section Business Analytics, Amsterdam Business School, University of Amsterdam, Amsterdam, The Netherlands

[3] Department of Public Health, School of Public Health, Bushehr University of Medical Science, Bushehr, Iran

[4] The Persian Gulf Tropical Medicine Research Center, The Persian Gulf Biomedical Science Research Institute, Bushehr University of Medical Science, Bushehr, Iran

three separate models predicted by five different ML algorithms: eXtreme Gradient Boosting (XGB), K-Nearest Neighbor (KNN), Random Forest (RF), bagged-CART (b-CART), and LogitBoost (LB). With the exception of KNN, the studied models show good predictive capabilities when evaluating relevant accuracy scores, such as area under the curve. By implementing an ensemble stacking approach (either a Neural Net or a General Linear Model) on top of the aforementioned ML algorithms the performance is further boosted. Ultimately, for the prediction of admission to the ICU, the ensemble stacking via a Neural Net achieved the best result with an accuracy of over 95%. For mortality at the ICU, the vanilla XGB performed slightly better (1% difference with the meta-model). To predict large length of stays both ensemble stacking approaches yield comparable results. Besides it direct implications for managing COVID-19 patients, the approach presented serves as an example how data can be employed in future pandemics or crises.

1 Introduction

The Intensive Care Unit known as ICU is a critical department in a hospital with special equipment and trained medical personnel for critically sick or injured individuals (Merriam-Webster, 2022). This unit is responsible to provide emergency care for those who require immediate treatment as to deal with life-threatening conditions. In times of crises like natural hazards and pandemics, which create an influx of patients, providing immediate health services for cases with critical situations becomes of paramount importance (Bohmer et al., 2020).

During the recent COVID-19 pandemic health system all over the world have been heavily burdened by the sudden influx of patients. Countries have been faced by numerous challenges while attempting to maintain the health system responsive and capable to provide essential health services (WHO Headquarters (HQ). 2021). The increasing number of hospital admission due to the vast prevalence of this virus has caused several problems for ICU wards such as overburdening of staff (Mehta et al., 2021) and shortages of medical resources, see for example Cohen and Y. van der M. Rodgers (2020). A recent cohort study in the USA revealed that strains on critical care capacity were associated with the increased number of ICU mortality for COVID-19 patients (Bravata et al., 2021). Therefore, there is impetus to reconsider and improve the management plan for the ICU.

A potentially beneficial source to assist healthcare providers is the vast amount of data captured. However, these huge volumes of medical data such as patient's characteristics, medication administration records, and genomic sequences, make it bewildering and perhaps impossible to make decisions as an individual. Fortunately, prediction models powered by Machine Learning (ML) are capable to learn and provide tangible insights and it is no surprise that it is reported that such models are becoming a necessity for the modern health systems (Beam and Kohane, 2018). ML can be seen as a subset of Artificial Intelligence (AI) which has the capability of emulating human intelligence (El Naqa and Murphy, 2015). The ML algorithms are classified into six types (Oladipupo, 2010): supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, transduction, and learning to learn. Supervised learning, the most common type of ML, attempts to predict the outcome for unseen data based on learning the mapping between the input variable and output variable

by means of train data (Cunningham et al., 2008). A common task in supervised learning is classification, in which the algorithm learns to classify an unknown object into one of a set of pre-determined classes (Carrizosa and Romero Morales, 2013), which is commonly employed in healthcare (Tomar and Agarwal, 2013). When the data is classified into two classes the problem is called binary classification; if the task is to classify the dataset into more than two classes the problem is referred to as multi-class classification.

In the light of the current COVID-19 pandemic, the integration of ML can be considered to cope with various challenges related to the management of healthcare resources, consideration of treatment plans, informing policies, and research challenges (Schaar et al., 2021). As mentioned, managing scarce resources is a key challenge in times of a pandemic, for example the distribution and production of face masks (Tirkolaee et al., 2022), but also efficient use of the limited capacity of the ICU. Our research question is whether ML-prediction models are capable to provide insights about the ICU. To do so we focus on a threefold of elements, which directly affect the ICU capacity (required number of beds): the admission to the ICU, the likelihood of an excessive length of stay (LOS), and the mortality.

In line with this research questions this study presents a comprehensive ML-based framework to predict three target variables: ICU admission, ICU mortality, and ICU LOS of hospitalized COVID-19 patients. For the prediction well-established ML algorithms are used. Using the advanced idea of ensemble stacking, which, by means of a meta-model, combines the separate ML approaches in a single model, one can benefit from the different predictive capabilities of each ML algorithm. Also, the research uncovers the important features to these three target variables, which from a medical point of view is of significant value. Finally, we point out that the same as for example the contemporary works (Goli and Malmir 2020; Alinaghian and Goli 2017) on disaster management this paper serves as an example how one can leverage ML in the case of a crisis situation.

The rest of the paper is organized as follows. Section 2 presents research works in the literature that relate to our study. In Sect. 3 the materials, including the dataset, and pre-processing actions on data and ML algorithms are described, which also provides the relevant features on which the ML algorithms will be trained. The ML prediction results are presented in Sect. 4. Finally, we discuss the results and conclude in Sect. 5 and Sect. 6 respectively.

2 Related works

There is a wide field of literature on ML in healthcare. We summarize relevant literature that focus on the ICU. For example there are various works on applying ML models on the ICU admission and mortality for hospitalized COVID-19 patients, see for example Altini et al. (2021), Campbell et al. (2021), Hou et al. (2021), Podder and Mondal (2020), Ryan et al. (2020) and Vaid et al. (2020). In more detail, a study in Spain (Aznar-Gimeno et al., 2021) in which the cohort information of 3623 patients is used to provide a decision-making tool to assist clinicians to estimate the risk of ICU admission or mortality. Chieregato et al. (2021) developed a hybrid machine learning/deep learning model to predict the need to ICU among COVID-19 patients. They used the data of 558 patients admitted to a hospital in Italy. In another study (Mahdavi et al., 2021), to predict the mortality prognosis which can help in declining the mortality rates, by applying invasive and noninvasive biomarker three ML models were presented. To predict the requirement to intensive care, Kim et al. (2020) used a nationwide cohort in South Korea including data of 100 hospitals. Applying ML approach, Izquierdo et al. (2020) find out that age, fever, and tachypnoea was the most parsimonious

predictor of ICU admission. The results of a study revealed that ensemble-based models perform better in predicting both ICU admission and mortality of COVID-19 cases (Subudhi et al., 2021). Hernández-Pereira et al. (2021) predicted the need of COVID-19 patients to regular hospital admission or intensive care unit admission using several ML algorithms. To predict the ICU admission in next 5 days, Famiglini et al. (2021) developed three ML models based on the complete blood count data.

Various studies focused on ML algorithms to only predict the fatality among COVID-19 individuals (Churpek et al., 2021; Kuno et al., 2022; Rozenbaum et al., 2021; Wanyan et al., 2021). In a research in Iran for predicting COVID-19 mortality, seven ML algorithms were used. Random forest showed a better performance in comparison to others (Moulaei et al., 2022). A multi-center cohort study was conducted to predict the ICU mortality of COVID-19 patients. The three ML models in this research presented acceptable and similar predictive performances (Lorenzoni et al., 2021). Parchure et al. (2020) used a ML-based approach for near-term COVID-19 in-hospital mortality. To predict the ICU outcome, Cunningham et al. (2008) applied an Explainable Boosting Machine approach. Elhazmi et al. (2022) applied conventional logistic regression and decision tree to predict 28-day ICU mortality for a cohort consisting of 14 hospitals in Saudi Arabia. To develop an in-hospital mortality score at admission for COVID-19 patients, Laino et al. (2022) used several supervised ML algorithms. For predicting the probability of death among inpatients COVID-19, Zarei et al. (2022) obtained the highest performance using the C5.0 decision tree algorithm.

For LOS, Ebinger et al. (2021) created three ML models on hospital days 1, 2 and 3 to classify COVID-19 patients' LOS into two classes. To predict the ICU admission, mortality, and survivors' LOS (Dan et al., 2020) developed three ML prediction models. They used support vector machine (SVM) algorithm for the three prediction models. Their models obtained acceptable performance. Based on the literature, only one paper considered predicting ICU admission, mortality, and LOS of COVID-19 patients' simultaneously, but it only made use of one specific ML algorithm.

The contribution of this study is to extend the studies above by developing a comprehensive framework to predict ICU admission, mortality and LOS of COVID-19 patients by applying and comparing several classical ML algorithms. As we find that the correlations between the models' predictions are low, we further leverage their individual predictive capabilities by integrating them in a stacking ensemble approach. Lastly, as also reported in literature, the outlined framework itself demonstrates how ML can be employed to swiftly retrieve valuable information for healthcare managers and practitioners, which insights specific to this case are summarized.

3 Materials and methods

The proposed framework of this study involving three main steps is illustrated in Fig. 1. The first phase demonstrates the database and the series of actions that are carried out, to prepare the data for modeling. These includes integration of datasets, data cleaning, dealing with missing values, balancing the dataset, and feature selection which will form the basis for the ML models to be applied. The output of this step is used as input for the second and third levels. The second module applies five different ML algorithms to make prediction and evaluate them based on the statistical index and ROC curves. In the third step, an ensemble approach, using one linear and one non-liner meta-learner algorithm predicts the outcomes. Lastly, the final results of the second and third phases compare and the best approach is
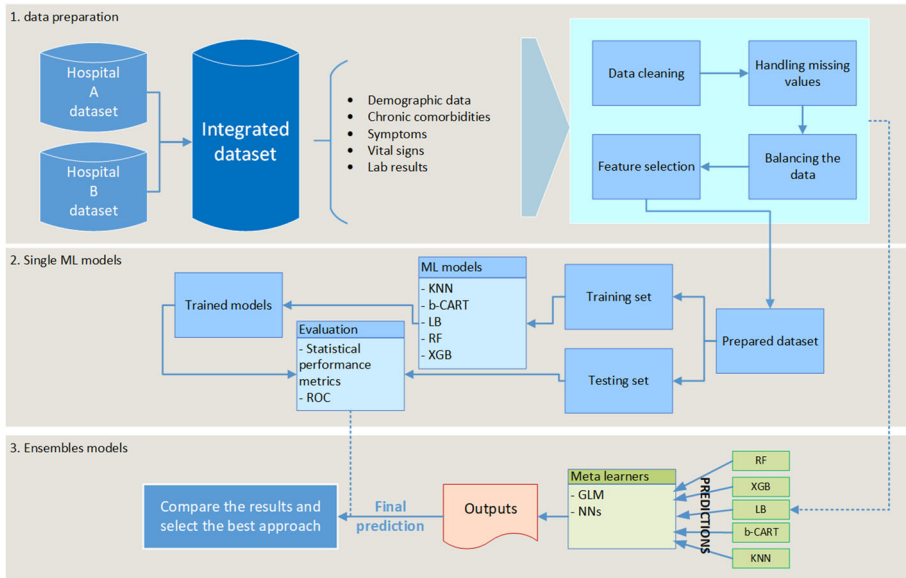
Fig. 1 Architecture of the framework of this study

selected. All of these steps are applied for the three ML models (ICU admission, mortality, and LOS) correspondingly. These processes are described in detail in the following sections. All analyses and modeling are done in R, which is an open-source programming language.

3.1 Study population

For this study, the data were collected from the files and electronic records of two Iranian local hospitals from September 7, 2020 to March 7, 2021 comprising six months. All included patients were admitted to ICU and confirmed with positive real time reverse transcriptase polymerase chain reaction (RT-PCR) test for COVID-19. An ethic approval was obtained from the Bushehr University of Medical Science, Iran. The collected information consists of demographic data, chronic comorbidities, symptoms, vital signs, and laboratory results at admission. The initial database contained more than 200 variables, from which only the primary information collected at admission and lab test were extracted as in accordance to the purpose of this study. This led to a final dataset with a total of 41 input variables and the three outcome variables. The outcome variables included ICU admission versus non-ICU admission, ICU patient's mortality versus discharge, and to make ICU LOS a classification problem whether it is under 7 days versus more than 7 days. In the latter case the threshold of 7 days is chosen to reflect a *below average* LOS or an *above average* LOS as the mean LOS was around 7 days.

The flowchart illustrating the case selection is presented in Fig. 2. Of the total 963 patients who tested positive for COVID-19, 956 were kept in the dataset, whereas 7 were excluded due to the incomplete medical records. From the remaining total of 956 patients, 844 were admitted to the HDU (High Dependency Unit) and 112 received ICU care. The ICU group included direct admissions, and a set of transferred cases from the HDU. Among the ICU patients, about 31% died and 77 patients were ultimately discharged from this unit. To
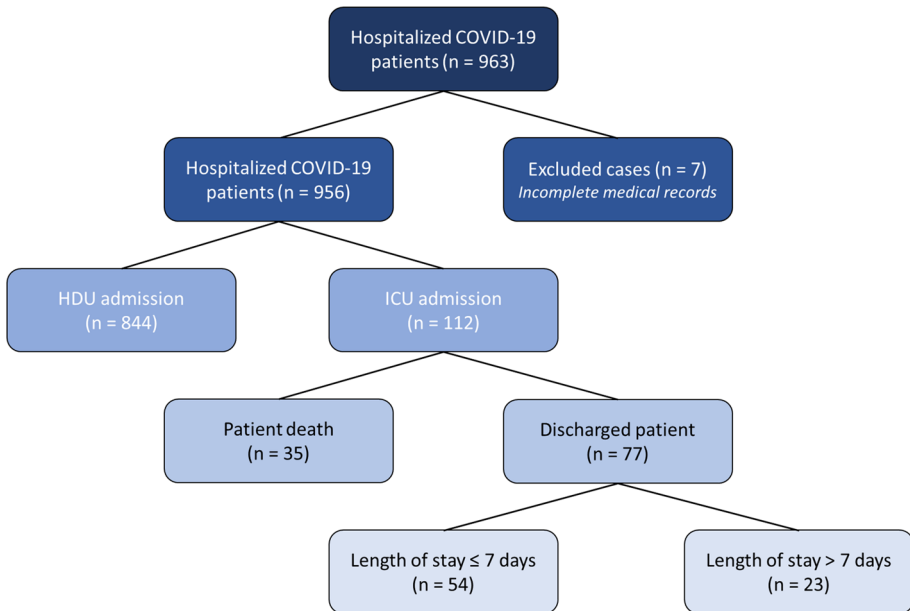
Fig. 2 The selection of cases of ICU admitted COVID-19 patients

determine the ICU LOS, the dead cases were excluded, such that 54 persons were hospitalized for 7 or fewer days, whereas 23 patients were hospitalized for a longer time period.

3.2 Missing values and data imputation

One of the common issues in medical data is the presence of missing values in independent variables (features), which if omitted, may cause a great reduction in sample size (Royston, 2004). In our dataset, as illustrated in Fig. 3. 6.8% of all observations that include 2669 fields relating to 40 different features were missing, and 36,527 fields, were fully completed. *LDH*, *Total bilirubin*, and *ESR* had the greatest number of empty fields and *breathing problem*, *cough*, and *systolic pressure* were the variables with the fewest missing values, respectively. Furthermore, *ARI (acute respiratory infection)* and *NCD (non-communicable diseases)* showed no blank fields at all. Fortunately, our three target variables, ICU admission, ICU deaths, and ICU LOS did not contain any missing values.

To decide how to deal with the variables with a high rate of empty fields medical specialist were consulted. These talks identified some variables with a high rate of empty fields as important (such as *ESR* and *LDH*), because they likely have a high impact on the target variables. Therefore, it was decided to keep all of them and—instead of deleting—to apply a suitable approach for data imputation.

To impute the data, the Multivariate Imputation by Chained Equations (MICE) algorithm also known as "fully conditional specification" in R was applied. This R package, imputes incomplete multivariate data by chained equations (Buuren and Groothuis-Oudshoorn, 2011). Single imputation methods, like use of mean and median and maximum likelihood methods have some limitation. The former, ignores uncertainty that may lead to excessively accurate results and the latter, is used for specific kind of models such as longitudinal or structural
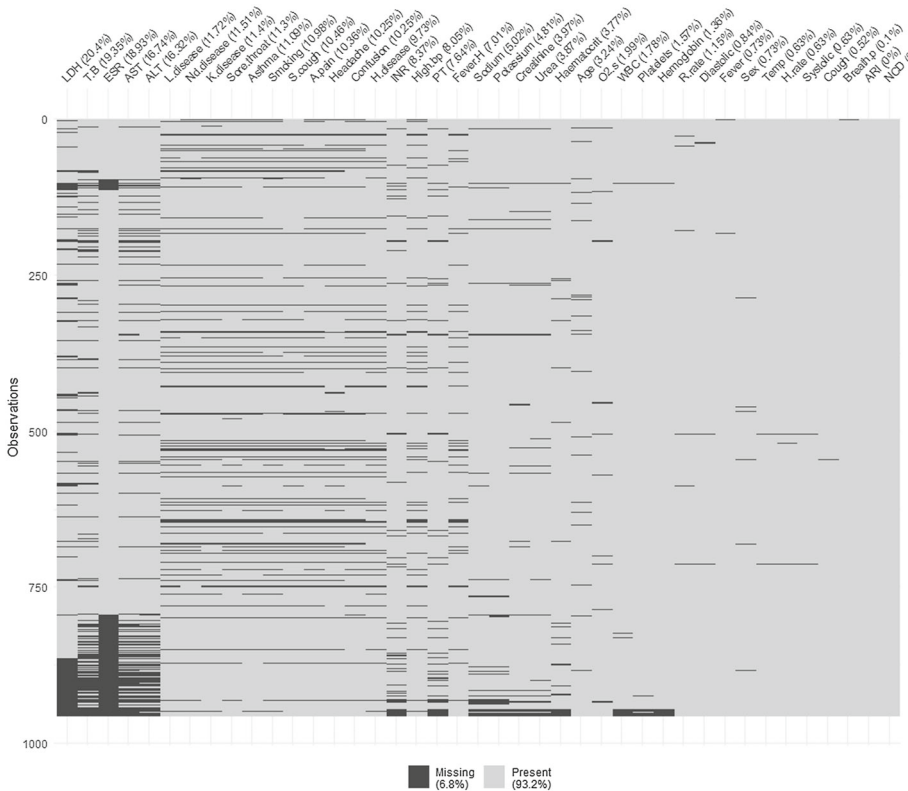
Fig. 3 Percentage of missing values of the dataset. Abbreviations: LDH (lactate dehydrogenase), T.B. (total bilirubin), ESR (erythrocyte sedimentation rate), AST (aspartate aminotransferase), ALT (alanine transaminase), L.disease (chronic lung disease), Nd.disease (chronic neurological disorder), K.disease (chronic kidney disease), S.cough (sputum cough), A.pain (abdominal pain), H.disease (heart disease), INR ((international normalized ratio), High.bp (high blood pressure), PT (prothrombin time), O2.s (O2 saturation), WBC (white blood cells) count, R.rate (respiratory rate), Diastolic (diastolic pressure), Temp (temperature), H.rate (heart rate), Systolic (systolic pressure), ARI (acute respiratory infection), NCD (non communicable diseases)

equation models that run under particular software (Azur et al., 2011). Compared to single imputation, the MICE algorithm has the benefit of considering uncertainty and multiple possible values in imputing missing data as reported by Zhang (2016).

To perform the MICE algorithm in this study, the number of multiple imputations was set to 5, which means for each missing value in the initial dataset there will be 5 probable values to be replaced. The selected imputation method for all variables was random forest with maximum iterations of 40. Kernel Density Estimation (KDE) used to estimate the probability density function of both initial and imputed data for variables. KDE is a widely used data smoothing technique which plots the data and creates distribution curves (Gramacki, 2018). Figure 4 illustrates the density plots of observed versus imputed data for 23 features out of 40. Generally, imputed values demonstrate acceptable distribution compared to observed values.
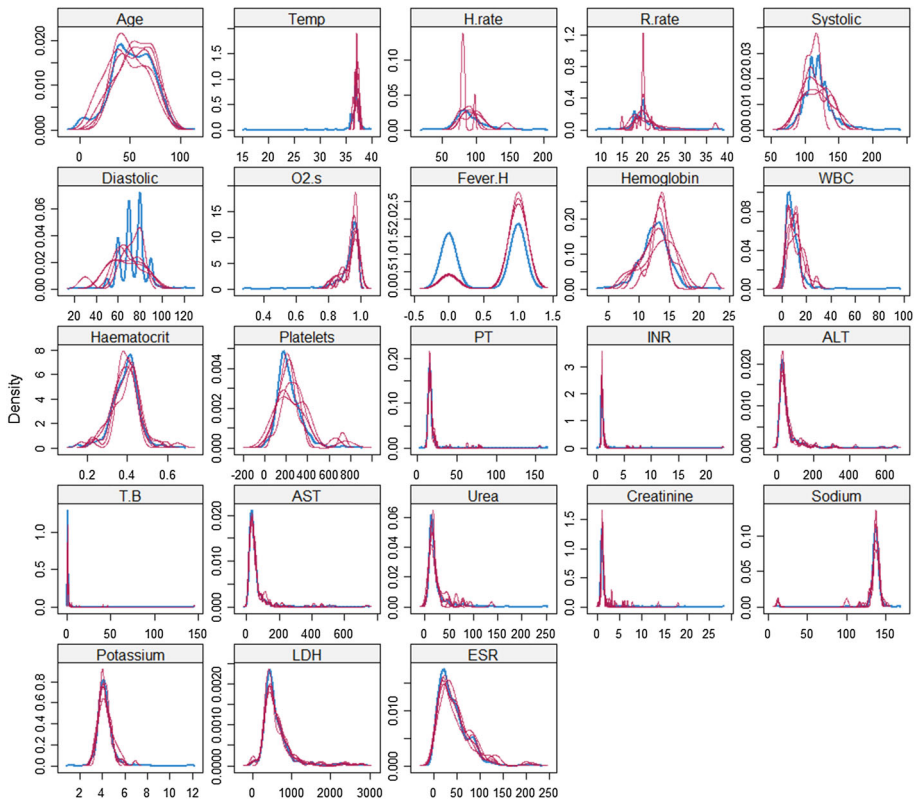
Fig. 4 Kernel Density Estimation of initial and imputed data for some of the variables. The red curves denote the imputed data distribution and the blue curves demonstrate the distribution of initial data. Abbreviations: Temp (temperature), H.rate (heart rate), R.rate (respiratory rate), Systolic (systolic pressure), Diastolic (diastolic pressure), O2.s (O2 saturation), Fever.H (history of fever), PT (prothrombin time), INR (international normalized ratio), ALT (alanine transaminase), LDH (lactate dehydrogenase), ESR (erythrocyte sedimentation rate)

3.3 Balancing the dataset

Another issue in datasets with the goal of classification is class imbalance (in the target variable), which hinders a ML algorithm to distinguish the relatively uncommon, but important class (Kotsiantis et al., 2005). Imbalances can manifest itself between classes, when one class has more examples than the other, or among some subsets of one class (Gu et al., 2008). In such situations, it is reported that classification ML models demonstrate poor performance and unreal predictions (Poolsawad et al., 2014). This problem origins from the assumptions of learning algorithms that consider accuracy (overall error) minimization as a goal in which the minority class contributes very little (Visa and Ralescu, 2005). From various domains of real-world datasets, medical data usually include a low number of positive or special cases against relatively many negatives.

In this research, the ICU admission variable, which is a binary variable was clearly imbalanced. This imbalance affected the prediction results badly and led the trained model to highly accurate results without considering the minority cases (admitted patients to ICU).

To deal with this issue, the ROSE (Random Over-Sampling Examples), an R package for binary classification problems, was used (Lunardon et al., 2014). ROSE is a synthetic data generation method which produces artificial data based on a bootstrap approach. The primarily percentage for the two classes of ICU admission were 11.8% and 88.1% for ICU admitted and non-admitted. After data balancing, the numbers changed to 52.2% and 47.8% respectively, which in turn will yield more accurate predictions.

3.4 Feature selection

To select the relevant attributes for each of the three ML models, the R package Boruta algorithm was applied (Kursa and Rudnicki, 2010). This method is an extension of the random forest algorithm by providing criteria for selection of important features (Kursa et al., 2010). For all models, the maximum random forest run was set to 500, and the *doTrace* which refers to the verbosity level was set to 2. The graph of variable importance for ICU admission is shown in Fig. 5 In this figure, the X axis represents the features and the Y axis indicates the importance of these attributes in predicting the target variable. The three blue box-plots are correspond to the minimum, average, and maximum of shadow variables. The irrelevant features are given the color red, whereas the green box-plots represent the features that are qualified as important. A yellow box-plot indicates that these variables are tentative as the algorithm cannot advise to include (*confirm*) or exclude the feature. For ICU admission, from

Fig. 5 The Boruta algorithm feature selection for ICU admission. Green boxes denote the confirmed features, yellow boxes represent the tentative variables, blue boxes illustrate the minimum, average, and maximum of shadow variables, and red boxes show the irrelevant features
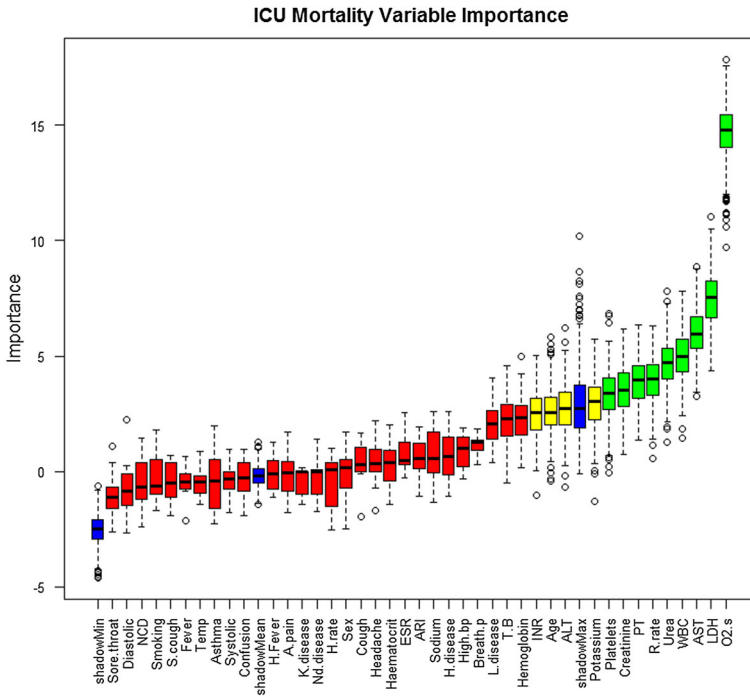
Fig. 6 The Boruta algorithm feature selection for ICU mortality. Green boxes denote the confirmed features, yellow boxes represent the tentative variables, blue boxes illustrate the minimum, average, and maximum of shadow variables, and red boxes show the irrelevant features

the total of 41 variables, 16 confirmed as accepted, 3 as tentative, and 22 as rejected. The top five features are in decreasing order of importance: *total bilirubin*, *INR*, *sodium, PT,* and *temperature*.

The top five relevant attributes to predict the ICU mortality based on the Boruta result are *O2 saturation, LDH, AST, WBC,* and *Urea* (Fig. 6). From all variables, 28 were considered unimportant, and 4 as tentative. Important variables related to predict of ICU LOS that are *hematocrit* and *ESR*, which are depicted in Fig. 7. *Total bilirubin* and *NCD* determined tentative, and the rest of the features were qualified unimportant.

3.5 Machine learning models

Three separate ML models were developed to predict three probable outcomes: (1) the need of ICU admission, (2) ICU mortality versus survival, and (3) LOS at the ICU for more, or less, than 7 days. According to feature selection results of the previous section, 27, 15, and 4 features were selected as input variables for our models respectively. To predict these three variables, we are employing five established, yet well-performing ML algorithms: RF, BL, b-CART, KNN, and XGB; each of them is briefly explained below. Furthermore, we provide the settings (hyperparameters) chosen as obtained by means of cross-validation on the train set; for more details see Sect. 4.
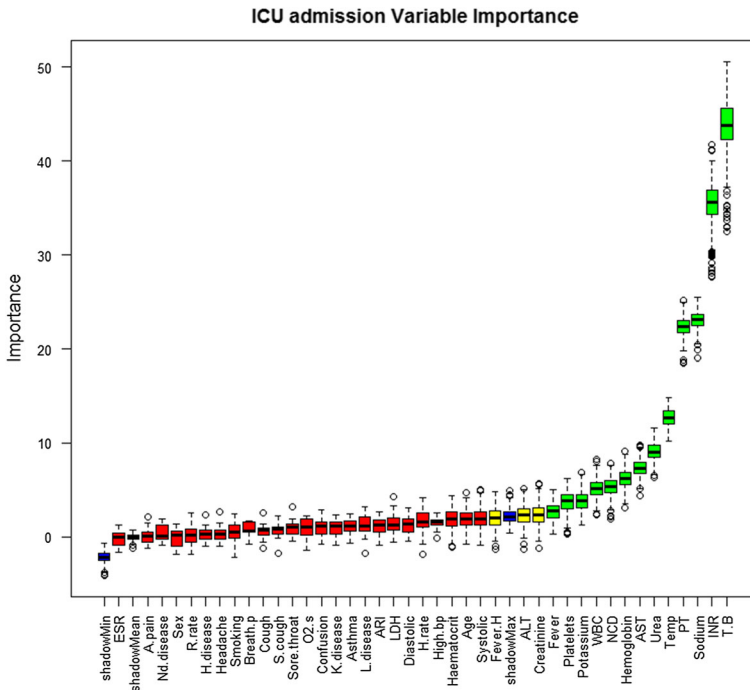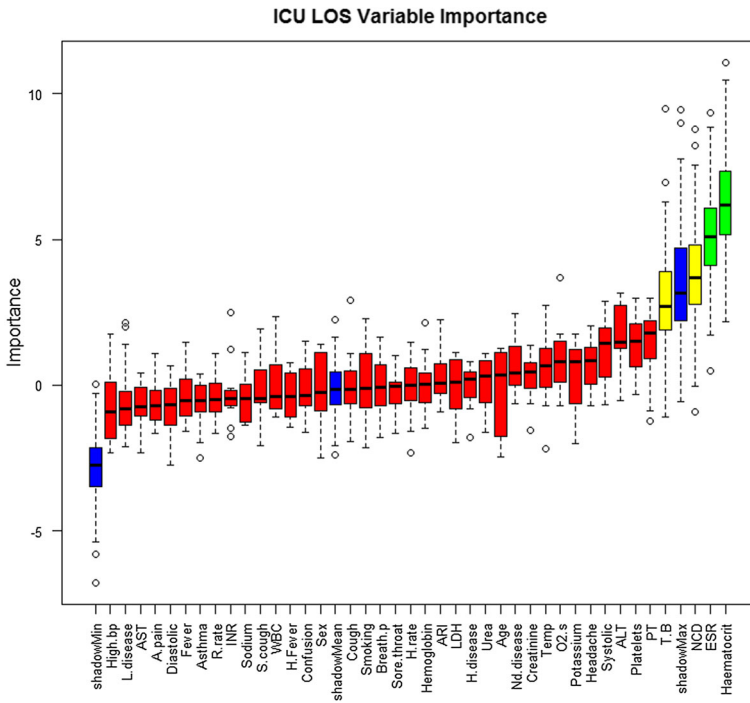
Fig. 7 The Boruta algorithm feature selection for ICU LOS. Green boxes denote the confirmed features, yellow boxes represent the tentative variables, blue boxes illustrate the minimum, average, and maximum of shadow variables, and red boxes show the irrelevant features

3.5.1 Random forest

The RF is a supervised Decision Tree based ML algorithm, which has the capability of coping with the overfitting problem (Breiman, 2001). This ensemble method constructs a multitude of decision trees on different samples and utilizes them for classifying an element based on the majority vote (Oshiro et al., 2012). Alongside making predictions, RF is capable of determining variable importance according to their impact on predicting the target variable (Boulesteix et al., 2012). To specify the best branch to split and thus the importance of a variable, the RF applies a splitting criterion, for example by using the *Gini impurity*. This index computes the overall probability of misclassifying at a node (Qi, 2012), and it is calculated as:

$$Gini = 1 - \sum_{i=1}^{c} p_i^2 \tag{1}$$

where p_i denotes the frequency of class i at a node and c represents the number of classes in the target variable. It ranges from 0 to 1. So, while one makes subsequential branching decisions one should opt to choose a split that lowers the weighted sum of the resulting indices the most. By continuing the process until a stopping criterion, e.g., the maximal number of data points at a node (terminal node size), a *tree* is formed. These trees are the basis for the random forest (RF) algorithm as it constitutes a *randomly* generated collection of trees.

To set the parameters for the RF algorithm, the number of trees to grow after each split (*ntree*) is 200 in this study. The final values for *mtry*, which refer to the number of variables randomly sampled as candidates at each split time, are 2, 2, and 6 for ICU admission, mortality, and LOS. The minimum terminal node size for all three prediction models is one, and there is no limitation on the maximum number of terminal nodes.

3.5.2 LogitBoost

Boosting algorithms include several weak learners, which will be combined to construct a final powerful learner. One of the popular boosting algorithms is LogitBoost (LB) proposed by Friedman et al. (2000). This method can be seen as the successor of the Adaboost algorithm which was sensitive to outliers and noise. Applying a binomial log-likelihood instead of an exponential loss function is brought up as a solution to this vulnerability, see (Kamarudin et al., 2017) for a discussion. LB consists of three main elements: (1) a multi-class logistic loss, (2) additive tree models, and (3) an optimization algorithm which minimizes the logistic loss (Sun et al., 2014). So, this algorithm minimizes the logistic loss over the training dataset of size *n*

$$Loss = -\sum_{i=1}^{n} \log\left(1 + e^{-2y_i F(x_i)}\right),\tag{2}$$

where F denotes the final classifier based on the features contained in the vector x_i and $y_i \in \{-1, 1\}$ is the set of labels (Karlos et al., 2015). Considering the model parameters, the final number of iterations for which boosting should be run for ICU admission, mortality, and LOS were set to 21, 31, and 11 respectively.

3.5.3 Bagged CART

Bootstrap aggregating, often referred to as bagging is a common ensemble method to reduce the problem of overfitting problem and thereby improving the accuracy on the test set (Breiman, 1996). Bagging can be applied to high-variance algorithms such as classification and regression trees (CART). The first step in the bagging algorithm is creating bootstrapped samples from the training dataset. Then, train, either classification or regression on each subset, and finally, aggregate the results by simply taking the average or majority vote in case of regression or classification (Polikar, 2006). In this study, the bagged CART was used for the three classification models. CART algorithm splits a node based on the Gini Index criterion (Rutkowski et al., 2014). The value of ensemble size (*nbagg*) for all prediction models was considered 25.

3.5.4 KNN

The k-nearest neighbor (KNN), firstly developed by Fix and Hodges (1989), is a non-parametric method that can be used for both classification and regression problems. Since this algorithm considers the whole dataset each prediction and does not require a specific training stage, it is called a *lazy* learner algorithm. The KNN classifies a new data point in the test set based on k points that are relatively close, so-called neighbours. Therefore one needs to introduce a concept of distance, which is readily incorporated in KNN algorithms

by relying on the Minkowski distance. The distance between two points, represented by the vectors x and x', is calculated as, where x_i denotes the i th element of a vector x:

$$d_{Minkowski} = \left(\sum_{i=1}^{n} |x_i - x'_i|^p \right)^{1/p} \tag{3}$$

So, p (a positive value) permits the calculation of different distance measures such as the standard Euclidean ($p = 2$) and Manhattan distance ($p = 1$). For more discussion on the choice of the distance measures, see Abu Alfeilat et al. (2019). We rely for this research on the standard. The final settings for k are 5, 5, and 7 – uneven numbers to break ties—for predicting ICU admission, mortality, and LOS, respectively.

3.5.5 Extreme gradient boosting

Extreme gradient boosting (XGB) is a celebrated ensemble ML algorithm based on the gradient boosted decision trees framework, which works more efficiently in terms of speed and performance compared to most ML approaches (Chen and Guestrin, 2016). In this algorithm, the objective function (measuring the model performance) consists of two parts: a loss function L and a regularization term Ω.

$$\mathcal{L}(\phi) = \sum_i L(\widehat{y_i}, y_i) + \Omega(F) \tag{4}$$

So when training, L evaluates the loss between a prediction $\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i)$ with f_k additive decision trees (which are found in successive and efficient manner by considering residuals) and y_i, while Ω avoids overfitting by regulating the model $F = \sum_{k=1}^{K} f_k(x_i)$. The hyperparameters of the XGB algorithms for the three prediction models are exhibited in Table 1. Maximum depth refers to the longest path between the root node and a leaf. Higher values of this parameter make the model more complex and may lead to overfitting. Eta, which lies within 0 and 1, controls the learning rate. Gamma determines the minimum loss reduction for making a split. Column sample by tree specifies the subsample ratio of columns when a new tree is constructed. Minimum child weight is the minimum sum of sample weight required in a child. Subsample denotes the ratio of the training instances (Chen and Guestrin, 2016).

To optimize the hyper parameters for all the algorithms (RF, LB, b-CART, XGB, and KNN) in the train set, the grid search method in Caret package (Kuhn, 2008) in R programming language were used.

4 Results

In this section we present our results. We first discuss the findings in our data, after which we apply the ML algorithms that are introduced in the previous section. Then we compare the models to conclude that there is potential to combine them by means of a meta-model, which results a so-called ensemble model.

4.1 Data description

All 41 predictor attributes can be classified into four groups: demographic information, symptoms, patient background, and lab results (Table 2). About 57% of all hospitalized patients belong to the age category of 19–60 years, which also has the highest rate compared to other categories among all ICU admitted and survived cases. For non-survived individuals,

Table 1 The XGB algorithm parameters setting

Prediction model	XGB parameter	Value
ICU admission	Maximum depth	3
	Eta	0.4
	Gamma	0
	Column sample by tree	0.6
	Minimum child weight	1
	Subsample	1
ICU mortality	Maximum depth	2
	Eta	0.3
	Gamma	0
	Column sample by tree	0.6
	Minimum child weight	1
	Subsample	0.5
ICU LOS	Maximum depth	3
	Eta	0.4
	Gamma	0
	Column sample by tree	0.8
	Minimum child weight	1
	Subsample	0.5

the category of age ≥ 60 consists 60% of all patients. Cough with a prevalence of 60% is one of the common symptoms among dead cases in the ICU. Breathing problem such as shortness of breath or short and rapid breathing at admission have the highest frequency for all admitted, ICU, and non-survived ICU patients. About 63% of all ICU admitted individuals experienced fever before visiting the hospital. On the other end, sputum cough and abdominal pain have the lowest occurrence.

In the case of patients' background diabetes is most prevalent as of the total of 956 hospitalized cases, 317 (33.15%) patients have diabetes. In the second place, high blood pressure with the occurrence of approximately 25% for all cases, 28.31% in all ICU patients, 26.92% in survived ICU, and 31.42% in dead ICU individuals received the highest rate. In the lab result category, the prothrombin time (PT) of 48.57% of non-survived ICU cases was more than 17 s. The total bilirubin of 30.08% of all ICU patients and 34.28% of non-survived ICU persons was more than 1 mg per deciliter (mg/dL). Among the 35 dead ICU cases, the AST lab results of 17 people were more than 54 units per liter of serum. For Creatinine factor, 512 of all patients showed values more than 1 mg/dL which its frequency for non-alive ICU people was 62.85%. The sodium level of 60% of non-survived ICU patients was less than equal to 137 milli-equivalents per liter (mEq/L). Lactic Acid Dehydrogenase (LDH) results of 331 of all cases indicated results of more than 600 units per liter (U/L). Among ICU patients, 55 cases out of 112 demonstrate Erythrocyte Sedimentation Rate (ESR) more than 42 mm per hour (mm/hr).

Table 2 Statistics of hospitalized patients' information

Variable		All patients (n = 956)	All ICU patients (n = 112)	Survived ICU patients (n = 77)	Non-survived ICU patients (n = 35)
Demographic	Sex (female)	464 (48.53%)	52 (46.01%)	35 (44.87%)	17 (51.42%)
	Sex (male)	492 (51.46%)	61 (53.98%)	43 (55.12%)	18 (48.57%)
	Age ≤ 18	52 (5.43%)	3 (2.65%)	2 (2.56%)	1 (2.85%)
	19–60	554 (57.94%)	66 (58.40%)	53 (67.94%)	13 (37.14%)
	Age > 60	350 (36.61%)	44 (38.93%)	23 (29.48%)	21 (60%)
Symptoms and clinical criteria at admission	History of Fever	413 (43.20%)	43 (38.05%)	29 (37.17%)	14 (40%)
	Cough	436 (45.60%)	59 (52.21%)	38 (48.71%)	21 (60%)
	Breathing problem (Shortness of breath or short and rapid breathing)	555 (58.05%)	69 (61.06%)	44 (56.41%)	25 (71.42%)
	Clinical suspicion to ARI	487 (50.94%)	52 (46.01%)	37 (47.43%)	15 (42.85%)
	Temperature ≤ 37	561 (58.68%)	66 (58.40%)	48 (61.53%)	18 (51.42%)
	Temperature > 37	395 (41.31%)	47 (41.59%)	30 (38.46%)	17 (48.57%)
	Heart rate ≤ 89	567 (59.30%)	60 (53.09%)	42 (53.84%)	18 (51.42%)
	Heart rate > 89	389 (40.69%)	53 (46.90%)	36 (46.15%)	17 (48.57%)
	Respiratory rate ≤ 20	738 (77.19%)	83 (73.45%)	60 (76.92%)	23 (65.71%)
	Respiratory rate > 20	218 (22.80%)	30 (26.54%)	18 (23.07%)	12 (34.28%)
	Systolic blood pressure ≤ 120	648 (67.78%)	75 (66.37%)	51 (65.38%)	24 (68.57%)
	Systolic blood pressure > 120	308 (32.21%)	38 (33.62%)	27 (34.61%)	11 (31.42%)
	Diastolic blood pressure ≤ 74	490 (51.25%)	56 (49.55%)	41 (52.56%)	15 (42.85%)
	Diastolic blood pressure > 74	466 (48.74%)	57 (50.44%)	37 (47.43%)	20 (57.14%)
	O2 saturation ≤ 0.9307	341 (35.66%)	49 (43.36%)	23 (29.48%)	26 (74.28%)
	O2 saturation > 0.9307	615 (64.33%)	64 (56.63%)	55 (70.51%)	9 (25.71%)
	History of fever	424 (44.35%)	44 (38.93%)	29 (37.17%)	15 (42.85%)
	Sputum cough	135 (14.12%)	18 (15.92%)	49 (62.82%)	8 (22.85%)
	Sore throat	87 (9.10%)	11 (9.73%)	7 (8.97%)	4 (11.42%)
	Confusion	167 (17.46%)	20 (17.69%)	13 (16.66%)	7 (20%)
	Headache	243 (25.41%)	30 (26.54%)	23 (29.48%)	7 (20%)
	Abdominal pain	104 (10.87%)	13 (11.50%)	11 (14.10%)	2 (5.71%)

Table 2 (continued)

Variable		All patients (n = 956)	All ICU patients (n = 112)	Survived ICU patients (n = 77)	Non-survived ICU patients (n = 35)
Patient background	Chronic heart disease	165 (17.25%)	23 (20.35%)	16 (20.51%)	7 (20%)
	High blood pressure	236 (24.68%)	32 (28.31%)	21 (26.92%)	11 (31.42%)
	Chronic lung disease	88 (9.20%)	18 (15.92%)	8 (10.25%)	10 (28.57%)
	Asthma	61 (6.38%)	9 (7.96%)	5 (6.41%)	4 (11.42%)
	Chronic kidney disease	73 (7.63%)	6 (5.30%)	5 (6.41%)	1 (2.85%)
	Chronic neurological disorder	50 (5.23%)	6 (5.30%)	5 (6.41%)	1 (2.85%)
	Smoking	96 (10%)	13 (11.50%)	9 (11.53%)	4 (11.42%)
	Diabetes	317 (33.15%)	36 (31.85%)	24 (30.76%)	12 (34.28%)
	Diabetes and thyroid	5 (0.52%)	1 (0.88%)	0	1 (2.85%)
	Thyroid	30 (3.13%)	2 (1.76%)	1 (1.28%)	1 (2.85%)
	Cerebral vascular accident	13 (1.35%)	3 (2.65%)	2 (2.56%)	1 (2.85%)
Lab results	Hemoglobin ≤ 13	521 (54.49%)	59 (52.21%)	39 (50%)	20 (57.14%)
	Hemoglobin > 13	435 (45.50%)	54 (47.78%)	39 (50%)	15 (42.85%)
	WBC count ≤ 9	571 (59.72%)	57 (50.44%)	42 (53.84%)	15 (42.85%)
	WBC count > 9	385 (40.27%)	56 (49.55%)	36 (46.15%)	20 (57.14%)
	Hematocrit ≤ 0.4	485 (50.73%)	60 (53.09%)	39 (50%)	21 (60%)
	Hematocrit > 0.4	471 (49.26%)	53 (46.90%)	39 (50%)	14 (40%)
	Platelets ≤ 200	448 (46.86%)	59 (52.21%)	39 (50%)	20 (57.14%)
	Platelets > 200	508 (53.13%)	54 (47.78%)	39 (50%)	15 (42.85%)
	PT ≤ 17	662 (69.24%)	74 (65.48%)	56 (71.79%)	18 (51.42%)
	PT > 17	294 (30.75%)	39 (34.51%)	22 (28.20%)	17 (48.57%)
	INR ≤ 1	309 (32.32%)	30 (26.54%)	25 (32.05%)	5 (14.28%)
	INR > 1	647 (67.67%)	83 (73.45%)	53 (67.94%)	30 (85.71%)
	ALT.SGPT ≤ 52	684 (71.54%)	86 (76.10%)	64 (82.05%)	22 (62.85%)
	ALT.SGPT > 52	272 (28.45%)	27 (23.89%)	14 (17.94%)	13 (37.14%)
	Total bilirubin ≤ 1	715 (74.79%)	79 (69.91%)	56 (71.79%)	23 (65.71%)
	Total bilirubin > 1	241 (25.20%)	34 (30.08%)	22 (28.20%)	12 (34.28%)
	AST.SGOT ≤ 54	688 (71.96%)	83 (73.45%)	65 (83.33%)	18 (51.42%)
	AST.SGOT > 54	268 (28.03%)	30 (26.54%)	13 (16.66%)	17 (48.57%)
	Urea ≤ 22	707 (73.95%)	77 (68.14%)	59 (75.64%)	18 (51.42%)
	Urea > 22	249 (26.04%)	36 (31.85%)	19 (24.35%)	17 (48.57%)

Table 2 (continued)

Variable	All patients (n = 956)	All ICU patients (n = 112)	Survived ICU patients (n = 77)	Non-survived ICU patients (n = 35)
Creatinine ≤ 1	444 (46.44%)	50 (44.24%)	37 (47.43%)	13 (37.14%)
Creatinine > 1	512 (53.55%)	63 (55.75%)	41 (52.56%)	22 (62.85%)
Sodium ≤ 137	473 (49.47%)	57 (50.44%)	36 (46.15%)	21 (60%)
Sodium > 137	483 (50.52%)	56 (49.55%)	42 (53.84%)	14 (40%)
LDH ≤ 600	625 (65.37%)	72 (63.71%)	58 (74.35%)	14 (40%)
LDH > 600	331 (34.62%)	41 (36.28%)	20 (25.64%)	21 (60%)
ESR ≤ 42	556 (58.15%)	58 (51.32%)	40 (51.28%)	18 (51.42%)
ESR > 42	400 (41.84%)	55 (48.67%)	38 (48.71%)	17 (48.57%)
Potassium ≤ 4	394 (41.21%)	39 (34.51%)	26 (33.33%)	13 (37.14%)
Potassium > 4	562 (58.78%)	74 (65.48%)	52 (66.66%)	22 (62.85%)

4.2 Prediction models

Using the data for prediction, we applied five ML algorithms. To do so the data set was split randomly, but balanced, in the ratio of 80:20 for predicting ICU admission, and 70:30 for ICU mortality and ICU LOS—we chose 70:30 split in the latter two cases to ensure sufficient data points in the test set. For model validation we applied ten-fold cross-validation.

The Receiver Operating Characteristic (ROC) curves and corresponding area under the curve (AUC) metric are used for comparing and evaluating the performance of ML algorithms for each target variable. Cohen's kappa, accuracy, sensitivity, and specificity are other measurements which are used in the assessment. Note that kappa is a statistical metric for categorical variables, which takes into account chance agreement; it is zero if the agreement coincide with random guessing, and one if there is perfect agreement; for more information see McHugh (2012). Accuracy refers to correct predictions, while sensitivity and specificity denote the rates of true positive and true negatives; they are defined as:

$$Accuracy = \frac{True\ positive\ predictions + True\ negative\ predictions}{All\ predictions}; \quad (5)$$

$$Sensitivity = \frac{True\ positive\ predictions}{(True\ positive\ predictions + False\ negative\ predictions)}; \quad (6)$$

$$Specificity = \frac{True\ negative\ predictions}{(True\ negative\ predictions + False\ positive\ predictions)}. \quad (7)$$

Based on these measures, one can consider so-called balanced accuracy, which is especially useful in the case when dealing with imbalanced classes – as in our case. It is defined as:

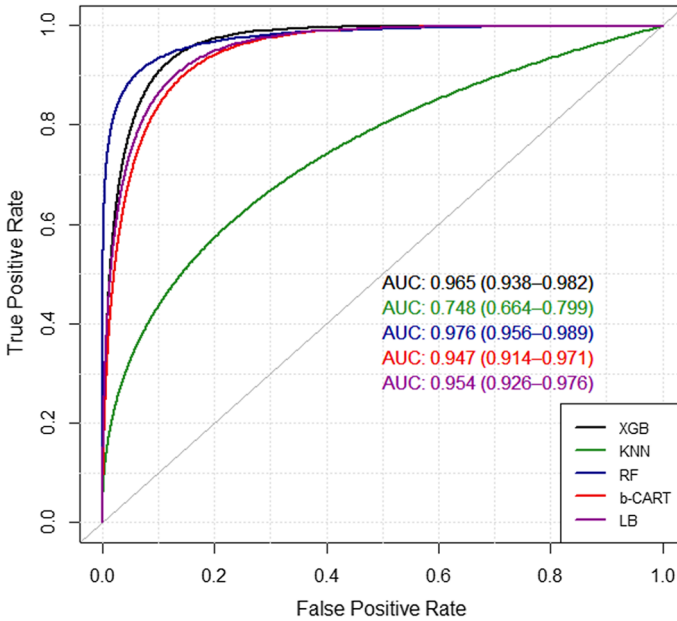$$Balanced\ accuracy = \frac{Sensitivity + Specificity}{2}, \quad (8)$$

Fig. 8 The ROC curves of five ML algorithms for ICU admission. Abbreviations: XGB (extreme gradient boosting), KNN (k-nearest neighbor), RF (random forest), b-CART (bagged CART), LB (LogitBoost)

Positive predictive value (PPV), also known as precision, and negative predictive value (NPV) are also calculated for assessing the predictive performance:

$$PPV = \frac{True\,positives}{True\,positives + False\,positives}, \quad \text{and}$$

$$NPV = \frac{True\,negatives}{True\,negatives + False\,negatives}. \tag{9}$$

Figure 8 demonstrates the ROC curves of five ML algorithms (b-CART, XGB, LB, KNN, and RF) for predicting ICU admission of COVID-19 patients. There is are slight differences between the AUC scores, except for KNN. RF with the AUC of 0.976 has the highest score followed by XGB and LB; the other metrics are shown in Table 3. We see that overall RF, and after that XGB, achieves the best performance. From the 19 selected independent variables to predict admission, the total bilirubin and INR were among the most important features, in line with the findings of Sect. 3.4.

The ROC curve of ICU mortality prediction is displayed in Fig. 9. The XGB obtained the highest AUC score (0.928), followed by KNN and RF with values of 0.917 and 0.868. Here LB was underperforming with only 0.746. In Table 4, we find that XGB also performs well in the other metrics.

For variable importance, out of the total 13 predictors that were selected for ICU mortality, O2 saturation and LDH were among five top attributes of all five algorithms. LDH with the value of 100 was the most important feature in XGB, KNN, RF, and LB algorithms, whereas in b-CART, LDH came second with an importance score of 78.06 and O2 saturation was most important. Other important features that were identified by the models were urea, PT,

Table 3 Performance metrics for ICU admission

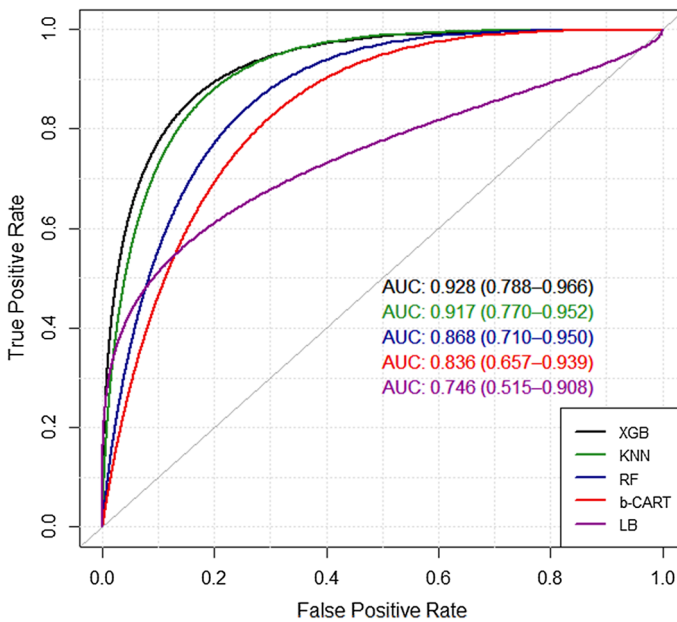| ICU admission								
Algorithm	Accuracy	95% CI	Sensitivity	Specificity	Kappa	Balanced accuracy	PPV	NPV
XGB	0.92	0.88–0.95	0.90	0.96	0.85	0.93	0.97	0.87
KNN	0.74	0.68–0.80	0.66	0.86	0.50	0.76	0.86	0.65
RF	0.93	0.89–0.96	0.91	0.96	0.87	0.94	0.97	0.89
CART	0.89	0.84–0.93	0.88	0.91	0.78	0.89	0.93	0.85
BLR	0.87	0.82–0.91	0.87	0.87	0.74	0.87	0.90	0.84

Fig. 9 The ROC curves of five ML algorithms for ICU mortality. Abbreviations: XGB (extreme gradient boosting), KNN (k-nearest neighbor), RF (random forest), b-CART (bagged CART), LB (LogitBoost)

Table 4 Performance metrics for ICU mortality

ICU mortality								
Algorithm	Accuracy	95% CI	Sensitivity	Specificity	Kappa	Balanced accuracy	PPV	NPV
XGB	0.78	0.61–0.91	0.90	0.58	0.51	0.74	0.79	0.77
KNN	0.78	0.61–0.91	0.95	0.50	0.49	0.72	0.76	0.85
RF	0.78	0.61–0.91	1.00	0.41	0.47	0.70	0.75	1.00
CART	0.72	0.54–0.86	0.90	0.41	0.35	0.66	0.73	0.71
BLR	0.75	0.57–0.88	0.95	0.41	0.41	0.68	0.74	0.83

INR and age. This interesting as the latter two were tentative features (see Sect. 3.4), but still therefore included in the model.

For predicting the ICU length of stay (LOS) Fig. 10 shows that XGB outperforms with an AUC score of 0.795, while RF is close with AUC of 0.778. Again, as for predicting admission KNN showed the worst performance. In terms of other performance metrics demonstrated in Table 5, we find that the b-CART algorithm generally provides better results. But, except for KNN the models have comparable scores, which is likely due to the fact that each model was fed with only four features.
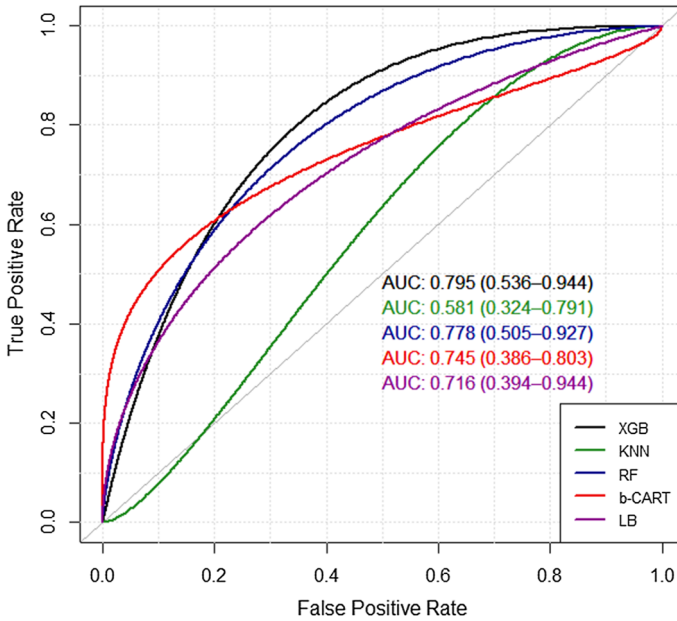

Fig. 10 The ROC curves of five ML algorithms for ICU LOS. Abbreviations: XGB (extreme gradient boosting), KNN (k-nearest neighbor), RF (random forest), b-CART (bagged CART), LB (LogitBoost)

Table 5 Performance metrics for ICU LOS

ICU LOS								
Algorithm	Accuracy	95% CI	Sensitivity	Specificity	Kappa	Balanced accuracy	PPV	NPV
XGB	0.78	0.56–0.92	0.88	0.40	0.31	0.64	0.84	0.50
KNN	0.69	0.47–0.86	0.83	0.20	0.03	0.51	0.78	0.25
RF	0.78	0.56–0.92	0.88	0.40	0.31	0.64	0.84	0.50
CART	0.82	0.61–0.95	0.88	0.60	0.48	0.74	0.88	0.60
BLR	0.69	0.47–0.86	0.72	0.60	0.26	0.66	0.86	0.37

4.3 Ensemble models

Ensemble algorithms are those learning methods which aim to construct a robust and more accurate prediction model by combining multiple learning algorithms (Rokach, 2010). Homogeneous and heterogeneous ensembles are two ways of integrating weak learners in ensemble learning techniques (Alazzam et al., 2017). Weak learners also known as base models are any ML algorithms which perform slightly better than random guessing. Bagging and boosting are two popular homogeneous ensemble learning, and stacking is one of the heterogeneous algorithm. In this study, stacking ensemble learning was applied to predict three intended outcomes (ICU admission, mortality, and LOS) and then the results compared with the best performed ML algorithm. In stacked generalization or stacking algorithm first, various learning algorithms do prediction, then their results are integrated through a combiner algorithm, which is another machine learning method.

The starting point of using an ensemble method for improving model performance is variation among the base models, i.e., low correlations. The idea is that the lower correlations between models the better the accuracy of an ensemble model.

For our case the correlations between the ICU admission prediction algorithms are displayed in Table 6. The correlation coefficients range between -1 and 1, where -1 denotes a perfect negative correlation and 1 indicates a perfect positive correlation, but of course a correlation value of 0 is desired. The highest correlation is found between the b-CART and XGB with a value of 0.5518, whereas KNN and LB show overall the most correlations close to zero. For ICU mortality, the correlations are shown in Table 7. We see there that KNN and b-CART have the least in common with other ML algorithms. Finally in Table 8, studying the correlations among the algorithms for ICU LOS, we see that KNN represents the lowest correlation with other algorithms. Of course, its performance was also considerably off compared to the other four ML algorithms.

Table 6 Base models correlation for ICU admission

	XGB	KNN	RF	b-CART	LB
XGB	1.0000000	0.08674282	0.5102666	0.55187687	0.22189048
KNN	0.08674282	1.0000000	0.3568629	− 0.15716049	0.14833669
RF	0.51026659	0.35686285	1.0000000	0.26039198	0.24939612
b-CART	0.55187687	− 0.15716049	0.2603920	1.0000000	− 0.07965935
LB	0.22189048	0.14833669	0.2493961	− 0.07965935	1.0000000

Table 7 Base models correlation for ICU mortality

	XGB	KNN	RF	b-CART	LB
XGB	1.0000000	0.4291930	0.5834555	0.3202589	0.7118382
KNN	0.4291930	1.0000000	0.1055946	0.2291481	0.1906466
RF	0.5834555	0.1055946	1.0000000	0.4477695	0.5931113
b-CART	0.3202589	0.2291481	0.4477695	1.0000000	0.3259707
LB	0.7118382	0.1906466	0.5931113	0.3259707	1.0000000

Table 8 Base models correlation for ICU LOS

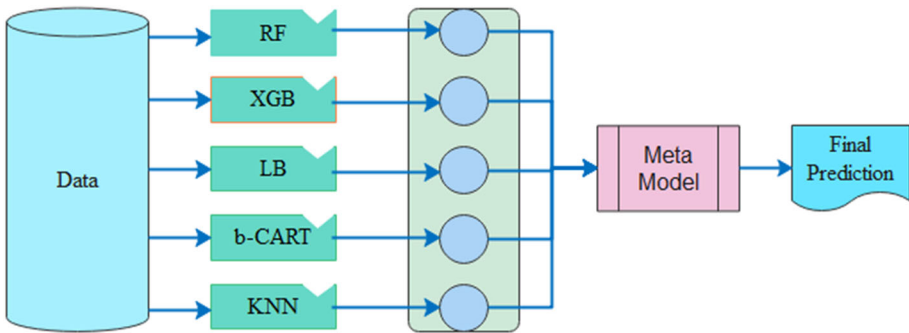|	XGB	KNN	RF	b-CART	LB
XGB	1.0000000	0.3316762	0.7209511	0.6036479	0.7071595
KNN	0.3316762	1.0000000	0.4432306	0.4737424	0.4432418
RF	0.7209511	0.4432306	1.0000000	0.8140235	0.5293713
b-CART	0.6036479	0.4737424	0.8140235	1.0000000	0.3637836
LB	0.7071595	0.4432418	0.5293713	0.3637836	1.0000000

Fig. 11 the overall concept of the ensemble method

To use the models in a combined fashion, one should stack them, which is illustrated in Fig. 11. The method consist of two phases. In our research, the first phase corresponds to the five ML algorithms (base models) that were trained to predict a target variable. The applied base algorithms are the same as the models used in Sect. 3.2. In the second phase, a so-called meta-model is introduced that integrates the predictions of these five base models, using the probability scores, into a single model, resulting a single probability score. To have a more comprehensive analysis, we consider as meta-model a Generalized Linear Model (GLM) and a single-layered Neural Network (NNs).

Doing this stacking by means of the two ensembling algorithms for each of the three target variables helps to answer whether the prediction performances can be improved. The results are presented in Table 9. A boldface value under the two metrics, accuracy and kappa, indicates that the corresponding prediction model performed the best compared with rival models. To predict the ICU admission, the ensemble model with the neural networks as the meta-model achieved the best result with an accuracy of 0.9577 (kappa of 0.9155). For ICU mortality, the XGB does a slightly better job. To predict the ICU LOS it seems that an ensemble model outperforms a single model, but the results are mixed as if considering accuracy the GLM as the meta-model wins, but if kappa is considered one should opt for the NN. These inconclusive results likely come from the fact that we deal with a limited dataset, both in number of features as well as data points.

Table 9 Performance metrics of best-model vs. ensemble models

Prediction model	Best-model vs. ensemble	Accuracy	Kappa
ICU admission	GLM (meta-model)	0.9570	0.9140
	NNs (meta-model)	**0.9577**	**0.9155**
	RF (best-applied model)	0.9481	0.8962
ICU mortality	GLM (meta-model)	0.8155	0.4901
	NNs (meta-model)	0.8111	0.5007
	XGB (best-applied model)	**0.8208**	**0.5254**
ICU LOS	GLM (meta-model)	**0.7114**	0.2431
	NNs (meta-model)	0.7040	**0.2794**
	XGB (best-applied model)	0.6750	0.2025

5 Discussion

Predicting the patient's disease course can help the resource allocation and planning in the hospital, which is especially important when there is a huge influx due to a pandemic. For that purpose, we developed three predictive models for ICU admission, mortality, and LOS by applying machine learning algorithms on hospitalized COVID-19 patients' data. Thereby, we show how data can be utilized to support physicians and healthcare staff in the early stage of COVID-19 patient admission at the hospital. Specifically, we determine the probability of admission to ICU, the mortality and whether a prolonged length of stay is likely.

As reported in our study, one of the difficulties with clinical data, perhaps even exacerbated in times of crises, is dealing with missing critical data such as lab results or accurate intake records. In our framework, instead of eliminating these values, an algorithm was applied to impute missing data. Another highlight is that, instead of relying on expert opinions, which might be unavailable for a variety of reasons, the algorithm provides the information and thus can accelerate the decision process. Moreover, the algorithm also selects which features (aspects) are important to consider. These features were similar with examined biomarkers related to the COVID-19 severity and mortality reported in literature, see for example Bousquet et al. (2020) and Yan et al. (2020).

According to the selected features for ICU admission and variable importance of the applied ML algorithms, total bilirubin is one of the key attributes in determining the requirement to the ICU. Several studies, such as Araç and Özel (2021), Liu (2020), and Roedl et al. (2021), confirm the impact of bilirubin on COVID-19 severity and mortality, specifically high levels of bilirubin are associated with higher probabilities of a severe disease course and mortality. Another prominent factor in the ICU admission prediction is creatinine. For example Ghosn et al. (2021) and Lowe et al. (2021) conclude that this feature is associated with admission, severity, mortality and LOS of hospitalized COVID-19 patients. Another important factor in our study to predict ICU admission and mortality is INR, which is confirmed in the systematic review of thirty-eight studies by A. Zinellu): Paliogiannis et al. (2021). These remarkable connections between our findings and the medical literature underpin the reliability and credibility of our model, and using ML in general.

For predicting the mortality of ICU patients, LDH is recognized as one of the important variables by our ML algorithms. The significance of this biomarker in the fatality rate of COVID-19 individuals is reported in for example Bousquet et al. (2020) and Yan et al.

(2020). As indicated in prior articles, e.g., Mejía et al. (2020) and Mansab et al. (2021), O2 saturation plays a prominent role whether a patient will survive and indeed it belongs to top fives of important variables of each ML algorithms that is applied. Also, Age is often recognized to be important to predict mortality and is backed by Bonanad et al. (2020), which performs a meta-analysis of COVID-19 cases from five different countries on the impact of age on the death rate. Finally, in our model to predict an extensive LOS at the ICU, the erythrocyte sedimentation rate (ESR) is crucial. A meta-analysis, which analyzed 16 studies, points out the association of inflammatory markers such as ESR with severity of COVID-19 cases (Zeng et al., 2020). In addition, the importance of Hematocrit (Hct) is supported by Kilercik et al. (2021), wherein much lower Hct values are observed among critical COVID-19 cases.

This research has several limitations. Firstly, the data stems from two local hospitals in Iran, which might limit the generalizability of results. Secondly, this study is limited to one wave of the COVID-19 pandemic. So, in the case of another wave, corresponding to for example a different strain of the virus, will likely yield slightly different results. We believe nevertheless, based on our discussion above, that the ML algorithms are capable to pick-up the important medical factors in such a new situation, because the framework and procedures followed are generic, i.e., not case dependent. Thirdly, considering the modeling, a limiting factor is the small-sized dataset, especially for predicting the LOS. If more data becomes available, it is likely that the predictive capability improves.

With more data at our disposal the models can be extended; currently the prediction revolves binary classification problems. But, with for example ICU admission or mortality one might also be interested in the time until such an event. So, in the case of a predicted ICU admission, when will this admission likely take place, or with a positive mortality outcome what is the most likely moment that the patient will decrease. For LOS one might be more interested to predict the actual duration or bed occupation than merely whether it will be more than 7 days. Note that in these extensions, instead of classification, one should consider it as regression problem. Finally, we scoped our study to predicting three variables by means of five established algorithms, a logical extension is to consider other ML algorithms and to predict other relevant variables, for example ones that relate to a treatment plan.

6 Conclusion

The main objective of this research is to provide a comprehensive approach for clinicians and managers to better manage scarce resources such as ICU beds, staff, and ventilators. Therefore, we propose a data-driven methodology using machine learning (ML) to predict ICU admission, mortality, and length of stay (LOS) of hospitalized COVID-19 patients. To alleviate the issue of missing values, and not to delete data, the MICE algorithm is applied. Then, because of the imbalanced classes in the datasets – which degrades the prediction performance – a synthetic data generation balancing method is used to create a balanced datasets. For these three outcome variables, potentially relevant features are selected by using the Boruta feature selection algorithm. Next, five different ML algorithms are applied, XGB, KNN, RF, b-CART, and BLR, which are all coded in R programming language. They show promising performance scores in terms of accuracy and AUC. Finally, in an attempt to further boost performance, an ensemble model is employed, which for predicting ICU admission and LOS outperforms relying on a single ML model and yields better accuracies

of 0.95 and 0.71 respectively. However, for ICU mortality, XGB with an accuracy of 0.82 outperforms ensembling.

Our research showcases how data, although the dataset is limited and incomplete, can be leveraged by means of ML to support decision makers in times of a healthcare crises, such as the COVID-19 pandemic, which centers this work. The fact that many of the key features of the prediction models coincide with the factors found in medical literature confirms the reliability and credibility of using our approach and ML in general. The models studied focus on determining whether the patient will be admitted to the ICU, will decease and whether a prolonged LOS is likely. Besides enriching this study with more data or repeating the study in other healthcare settings (different hospital, another COVID-19 wave, new virus) or for other variables, predicting the actual timings of those events is a logical starting point for further research.

Data availability Due to the sensitive nature of data used in this study, the hospital authority was assured raw data should remain confidential and should not be shared.

Declarations

Conflict of interests The authors declare that they have no conflict of interest.

Ethical approval This study was approved by the ethical committee of Bushehr University of Medical Science in Iran. The study was conducted in accordance with the ethical standards of the Helsinki declaration.

References

Abu Alfeilat, H. A. et al. (2019). Effects of distance measure choice on K-nearest neighbor classifier performance: A review. *Big Data, 7*(4), 221–248. https://doi.org/10.1089/big.2018.0175

Alazzam, I., Alsmadi, I., & Akour, M. (2017). Software fault proneness prediction: A comparative study between bagging, boosting, and stacking ensemble and base learner methods. *International Journal of Data Analysis Techniques and Strategies, 9*(1), 1. https://doi.org/10.1504/IJDATS.2017.10003991

Alinaghian, M., & Goli, A. (2017). Location, allocation and routing of temporary health centers in rural areas in crisis, solved by improved harmony search algorithm. *International Journal of Computational Intelligence Systems, 10*(1), 894. https://doi.org/10.2991/ijcis.2017.10.1.60

Altini, N. et al. (2021). Predictive machine learning models and survival analysis for COVID-19 prognosis based on hematochemical parameters. *Sensors, 21*(24), 8503. https://doi.org/10.3390/s21248503

Araç, S., & Özel, M. (2021). A new parameter for predict the clinical outcome of patients with COVID-19 pneumonia: The direct/total bilirubin ratio. *International Journal of Clinical Practice*. https://doi.org/10.1111/ijcp.14557

Aznar-Gimeno, R. et al. (2021). A clinical decision web to predict ICU admission or death for patients hospitalised with COVID-19 using machine learning algorithms. *International Journal of Environmental Research and Public Health, 18*(16), 8677. https://doi.org/10.3390/ijerph18168677

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research, 20*(1), 40–49. https://doi.org/10.1002/mpr.329

Beam, A. L., & Kohane, I. S. (2018). Big Data and Machine Learning in Health Care. *JAMA, 319*(13), 1317. https://doi.org/10.1001/jama.2017.18391

Bohmer, T., Pisano, R. M. J., Sadun, G. P., & Tsai, R. (2020). How hospitals can manage supply chain shortages as demand surges. *Harvard Business Review*. https://hbr.org/2020/04/how-hospitals-can-manage-supply-shortages-as-demand-surges.

Bonanad, C. et al. (2020). The effect of age on mortality in patients with COVID-19: A meta-analysis with 611,583 subjects. *Journal of the American Medical Directors Association, 21*(7), 915–918. https://doi.org/10.1016/j.jamda.2020.05.045

Boulesteix, A.-L., Janitza, S., Kruppa, J., & König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2*(6), 493–507. https://doi.org/10.1002/widm.1072

Bousquet, G. et al. (2020). ADL-dependency, D-Dimers, LDH and absence of anticoagulation are independently associated with one-month mortality in older inpatients with Covid-19. *Aging Albany NY, 12*(12), 11306–11313. https://doi.org/10.18632/aging.103583

Bravata, D. M. et al. (2021). Association of Intensive Care Unit Patient Load and demand with mortality rates in US Department of Veterans Affairs Hospitals During the COVID-19 pandemic. *JAMA Network Open, 4*(1), e2034266. https://doi.org/10.1001/jamanetworkopen.2020.34266

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Campbell, T. W. et al. (2021). Predicting prognosis in COVID-19 patients using machine learning and readily available clinical data. *International Journal of Medical Informatics, 155*, 104594. https://doi.org/10.1016/j.ijmedinf.2021.104594

Carrizosa, E., & Romero Morales, D. (2013). Supervised classification and mathematical optimization. *Computers & Operations Research, 40*(1), 150–165. https://doi.org/10.1016/j.cor.2012.05.015

Chen, T., & Guestrin, C. (2016). XGBoost. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785.

Chieregato, M. et al. (2021). A hybrid machine learning/deep learning COVID-19 severity predictive model from CT images and clinical data. May 2021 (**online**). Available: http://arxiv.org/abs/2105.06141.

Churpek, M. M. et al. (2021). Machine learning prediction of death in critically Ill patients with coronavirus disease 2019. *Critical Care Explorations, 3*(8), e0515. https://doi.org/10.1097/CCE.0000000000000515

Cohen, J., & van der Rodgers, Y. M. (2020). Contributing factors to personal protective equipment shortages during the COVID-19 pandemic. *Preventive Medicine (Baltim),*https://doi.org/10.1016/j.ypmed.2020.106263

Cunningham, P., Cord, M., & Delany, S. J. (2008) Supervised Learning. In *Machine Learning Techniques for Multimedia*, Springer: Berlin, pp. 21–49.

Dan, T. et al. (2020). Machine learning to predict ICU admission, ICU mortality and survivors' length of stay among COVID-19 patients: toward optimal allocation of ICU resources. In *2020 IEEE international conference on bioinformatics and biomedicine (BIBM)* (pp. 555–561). https://doi.org/10.1109/BIBM49941.2020.9313292.

Ebinger, J. et al. (2021). A machine learning algorithm predicts duration of hospitalization in COVID-19 patients. *Intelligence-Based Medicine,* https://doi.org/10.1016/j.ibmed.2021.100035

Elhazmi, A. et al. (2022). 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, 15*(7), 826–834. https://doi.org/10.1016/j.jiph.2022.06.008

Famiglini, L., Bini, G., Carobene, A., Campagner, A., & Cabitza, F. (2021). Prediction of ICU admission for COVID-19 patients: A machine learning approach based on complete blood count data. In *2021 IEEE 34th international symposium on computer-based medical systems (CBMS)* (pp. 160–165). https://doi.org/10.1109/CBMS52027.2021.00065.

Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. Nonparametric discrimination: consistency properties. *International Statistical Review, 57*(3), 238. https://doi.org/10.2307/1403797

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *Annals of Statistics*. https://doi.org/10.1214/aos/1016218223

Ghosn, M. et al. (2021). Severe acute kidney injury in critically Ill patients with COVID-19 admitted to ICU: Incidence, risk factors, and outcomes. *Journal of Clinical Medicine, 10*(6), 1217. https://doi.org/10.3390/jcm10061217

Goli, A., & Malmir, B. (2020). A covering tour approach for disaster relief locating and routing with fuzzy demand. *International Journal of Intelligent Transportation Systems Research, 18*(1), 140–152. https://doi.org/10.1007/s13177-019-00185-2

Gramacki, A. (2018). *Nonparametric kernel density estimation and its computational aspects* (Vol. 37). Berlin: Springer.

Gu, Q., Cai, Z., Zhu, L., & Huang, B. (2008). Data mining on imbalanced data sets. In *2008 international conference on advanced computer theory and engineering* (pp. 1020–1024) https://doi.org/10.1109/ICACTE.2008.26.

WHO Headquarters (HQ). (2021) Second round of the national pulse survey on continuity of essential health services during the COVID-19 pandemic: January–March 2021 (**online**). Available: https://www.who.int/publications/i/item/WHO-2019-nCoV-EHS-continuity-survey-2021.1.

Hernández-Pereira, E., Fontenla-Romero, O., Bolón-Canedo, V., Cancela-Barizo, B., Guijarro-Berdiñas, B., & Alonso-Betanzos, A. (2021). Machine learning techniques to predict different levels of hospital care of CoVid-19. *Applied Intelligence*. https://doi.org/10.1007/s10489-021-02743-2

Hou, W., Zhao, Z., Chen, A., Li, H., & Duong, T. Q. (2021). Machining learning predicts the need for escalated care and mortality in COVID-19 patients from clinical variables. *International Journal of Medical Sciences, 18*(8), 1739–1745. https://doi.org/10.7150/ijms.51235

Izquierdo, J. L., Ancochea, J., & Soriano, J. B. (2020). Clinical characteristics and prognostic factors for intensive care unit admission of patients with COVID-19: Retrospective study using machine learning and natural language processing. *Journal of Medical Internet Research, 22*(10), e21801. https://doi.org/10.2196/21801

Kamarudin, M. H., Maple, C., Watson, T., & Safa, N. S. (2017). A LogitBoost-based algorithm for detecting known and unknown web attacks. *IEEE Access, 5*, 26190–26200. https://doi.org/10.1109/ACCESS.2017.2766844

Karlos, S., Fazakis, N., Kotsiantis, S., & Sgarbas, K. (2015). Self-train logitboost for semi-supervised learning. In Iliadis, L., & Jayne, C. (Eds.), Engineering applications of neural networks. EANN 2015. Communications in Computer and Information Science, vol 517. Springer, Cham. https://doi.org/10.1007/978-3-319-23983-5_14

Kilercik, M., Demirelce, Ö., Serdar, M. A., Mikailova, P., & Serteser, M. (2021). A new haematocytometric index: Predicting severity and mortality risk value in COVID-19 patients. *PLoS ONE, 16*(8), e 0254073. https://doi.org/10.1371/journal.pone.0254073

Kim, H.-J. et al. (2020). An easy-to-use machine learning model to predict the prognosis of patients With COVID-19: Retrospective cohort study. *Journal of Medical Internet Research, 22*(11), e24225. https://doi.org/10.2196/24225

Kotsiantis, S., Dimitris, K., & Panayiotis, P. (2005). Handling imbalanced datasets: A review. *GESTS International Transaction on Computer Science and Engineering, 30*, 25–36.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v028.i05

Kuno, T., Sahashi, Y., Kawahito, S., Takahashi, M., Iwagami, M., & Egorova, N. N. (2022). Prediction of in-hospital mortality with machine learning for COVID-19 patients treated with steroid and remdesivir. *Journal of Medical Virology, 94*(3), 958–964. https://doi.org/10.1002/jmv.27393

Kursa, M. B., Jankowski, A., & Rudnicki, W. R. (2010). Boruta—A system for feature selection. *Fundam. Informaticae, 101*(4), 271–285. https://doi.org/10.3233/FI-2010-288

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v036.i11

Laino, M. E. et al. (2022). An individualized algorithm to predict mortality in COVID-19 pneumonia: A machine learning based study. *Archives of Medical Science, 18*(3), 587–595. https://doi.org/10.5114/aoms/144980

Liu, Z. et al. (2020). Bilirubin levels as potential indicators of disease severity in coronavirus disease patients: A retrospective cohort study. *Frontiers in Medicine*. https://doi.org/10.3389/fmed.2020.598870

Lorenzoni, G. et al. (2021). COVID-19 ICU mortality prediction: A machine learning approach using Super-Learner algorithm. *Journal of Anesthesia, Analgesia and Critical Care, 1*(1), 3. https://doi.org/10.1186/s44158-021-00002-x

Lowe, R. et al. (2021). Clinical characteristics and outcome of critically ill COVID-19 patients with acute kidney injury: A single centre cohort study. *BMC Nephrology, 22*(1), 92. https://doi.org/10.1186/s12882-021-02296-z

Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: A package for binary imbalanced learning. *The R Journal, 6*(1), 79. https://doi.org/10.32614/RJ-2014-008

Mahdavi, M. et al. (2021). A machine learning based exploration of COVID-19 mortality risk. *PLoS ONE, 16*(7), e0252384. https://doi.org/10.1371/journal.pone.0252384

Mansab, F., Donnelly, H., Kussner, A., Neil, J., Bhatti, S., & Goyal, D. K. (2021). Oxygen and mortality in COVID-19 pneumonia: A comparative analysis of supplemental oxygen policies and health outcomes across 26 countries. *Frontiers in Public Health*. https://doi.org/10.3389/fpubh.2021.580585

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Mehta, S. et al. (2021). COVID-19: A heavy toll on health-care workers. *The Lancet Respiratory Medicine, 9*(3), 226–228. https://doi.org/10.1016/S2213-2600(21)00068-0

Mejía, F. et al. (2020). Oxygen saturation as a predictor of mortality in hospitalized adult patients with COVID-19 in a public hospital in Lima, Peru. *PLoS ONE, 15*(12), e0244171. https://doi.org/10.1371/journal.pone.0244171

Merriam-Webster. (2022, n.d.). Intensive care unit. *Merriam-Webster.com dictionary*, 2022. https://www.merriam-webster.com/dictionary/intensive care unit.

Moulaei, K., Shanbehzadeh, M., Mohammadi-Taghiabad, Z., & Kazemi-Arpanahi, H. (2022). Comparing machine learning algorithms for predicting COVID-19 mortality. *BMC Medical Informatics and Decision Making, 22*(1), 2. https://doi.org/10.1186/s12911-021-01742-0

El Naqa, I., & Murphy, M. J. (2015). What is machine learning? In *Machine learning in radiation oncology*. Cham: Springer (pp. 3–11).

Oladipupo, T. (2010). Types of machine learning algorithms. In *New advances in machine learning*, InTech.

Oshiro, T. M., Perez, S., & Baranauskas, J. A. (2012). How many trees in a random forest? In P. Perner (Ed.), *Machine learning and data mining in pattern recognition* (pp. 154–168). Berlin: Springer.

Parchure, P. et al. (2020). Development and validation of a machine learning-based prediction model for near-term in-hospital mortality among patients with COVID-19. *BMJ Supportive & Palliative Care*, p. bmjspcare-2020-002602, Sept. 2020, https://doi.org/10.1136/bmjspcare-2020-002602.

Podder, P., & Mondal, M. R. H. (2020). Machine learning to predict COVID-19 and ICU requirement. In *2020 11th international conference on electrical and computer engineering* (*ICECE*) (pp. 483–486). https://doi.org/10.1109/ICECE51571.2020.9393123.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine, 6*(3), 21–45. https://doi.org/10.1109/MCAS.2006.1688199

Poolsawad, N., Kambhampati, C., & Cleland, J. G. (2014). Balancing class for performance of classification with a clinical dataset. In *Proceedings of the world congress on engineering*.

Qi, Y. (2012). Random forest for bioinformatics. In *Ensemble machine learning*. Boston, MA: Springer (pp. 307–323).

Roedl, K. et al. (2021). Severe liver dysfunction complicating course of COVID-19 in the critically ill: Multi-factorial cause or direct viral effect? *Annals of Intensive Care, 11*(1), 44. https://doi.org/10.1186/s13613-021-00835-3

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review, 33*(1–2), 1–39. https://doi.org/10.1007/s10462-009-9124-7

Royston, P. (2004). Multiple Imputation of Missing Values. *The Stata Journal: Promoting Communications on Statistics and Stata, 4*(3), 227–241. https://doi.org/10.1177/1536867X0400400301

Rozenbaum, D., Shreve, J., Radakovich, N., Duggal, A., Jehi, L., & Nazha, A. (2021). Personalized prediction of hospital mortality in COVID-19—positive patients. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes, 5*(4), 795–801. https://doi.org/10.1016/j.mayocpiqo.2021.05.001

Rutkowski, L., Jaworski, M., Pietruczuk, L., & Duda, P. (2014). The CART decision tree for mining data streams. *Information Sciences (NY), 266*, 1–15. https://doi.org/10.1016/j.ins.2013.12.060

Ryan, L. et al. (2020). Mortality prediction model for the triage of COVID-19, pneumonia, and mechanically ventilated ICU patients: a retrospective study. *Annals of Medicine and Surgery, 59*, 207–216. https://doi.org/10.1016/j.amsu.2020.09.044

Subudhi, S. et al. (2021). Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. *Npj Digital Medicine, 4*(1), 87. https://doi.org/10.1038/s41746-021-00456-x

Sun, P., Reid, M. D., & Zhou, J. (2014). An improved multiclass LogitBoost using adaptive-one-vs-one. *Machine Learning, 97*(3), 295–326. https://doi.org/10.1007/s10994-014-5434-3

Tirkolaee, E. B., Goli, A., Ghasemi, P., & Goodarzian, F. (2022). Designing a sustainable closed-loop supply chain network of face masks during the COVID-19 pandemic: Pareto-based algorithms". *Journal of Cleaner Production, 333*, 130056. https://doi.org/10.1016/j.jclepro.2021.130056

Tomar, D., & Agarwal, S. (2013). A survey on data mining approaches for healthcare. *International Journal of Bio-Science and Bio-Technology, 5*(5), 241–266. https://doi.org/10.14257/ijbsbt.2013.5.5.25

Vaid, A. et al. (2020). Machine learning to predict mortality and critical events in a cohort of patients with COVID-19 in New York City: Model development and validation. *Journal of Medical Internet Research, 22*(11), e24018. https://doi.org/10.2196/24018

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v045.i03

van der Schaar, M. et al. (2021). How artificial intelligence and machine learning can help healthcare systems respond to COVID-19. *Machine Learning, 110*(1), 1–14. https://doi.org/10.1007/s10994-020-05928-x

Visa, S., & Ralescu, A. (2005). Issues in mining imbalanced data sets—A review paper. In *Proceedings of the sixteen midwest artificial intelligence and cognitive science conference* (pp. 67–73).

Wanyan, T. et al. (2021). Relational learning improves prediction of mortality in COVID-19 in the intensive care unit. *IEEE Transactions on Big Data, 7*(1), 38–44. https://doi.org/10.1109/TBDATA.2020.3048644

Yan, L. et al. (2020). An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence, 2*(5), 283–288. https://doi.org/10.1038/s42256-020-0180-7

Zarei, J., Jamshidnezhad, A., Haddadzadeh Shoushtari, M., Mohammad Hadianfard, A., Cheraghi, M., & Sheikhtaheri, A. (2022). Machine learning models to predict in-hospital mortality among inpatients with COVID-19: Underestimation and overestimation bias analysis in subgroup populations. *Journal of Healthcare Engineering*. https://doi.org/10.1155/2022/1644910

Zeng, F. et al. (2020). Association of inflammatory markers with the severity of COVID-19: A meta-analysis. *International Journal of Infectious Diseases, 96*, 467–474. https://doi.org/10.1016/j.ijid.2020.05.055

Zhang, Z. (2016). Multiple imputation with multivariate imputation by chained equation (MICE) package. *Annals of Translational Medicine , 4*(2) (**online**). Available: https://atm.amegroups.com/article/view/8847.

Zinellu, A., Paliogiannis, P., Carru, C., & Mangoni, A. A. (2021). INR and COVID-19 severity and mortality: A systematic review with meta-analysis and meta-regression. *Advances in Medical Sciences, 66*(2), 372–380. https://doi.org/10.1016/j.advms.2021.07.009