Journal of International Technology and Information Management

Manuscript 1541

Short-term prediction of ICU admission for COVID-19 inpatients

Yoon Sang Lee

Riyaz T. Sikora

Short-term prediction of ICU admission for COVID-19 inpatients

Yoon Sang Lee
(Columbus State University)

Riyaz T. Sikora
(University of Texas at Arlington)

ABSTRACT

Since the COVID-19 outbreak, many hospitals suffered from a surge of some high-risk inpatients needing to be admitted to the ICU. In this study, we propose a method predicting the likelihood of COVID-19 inpatients' admission to the ICU within a time frame of 12 hours. Four steps, the Bayesian Ridge Regression-based missing value imputation, the synthesis of training samples by the combination of two rows (the first and another row) of each patient, customized oversampling, and XGBoost classifier, are used for the proposed method. In the experiment, the AUC-ROC and F-score of our method is compared with those of other methods using various imputation techniques and classifiers. Our method achieves the best performance among the methods.

Keywords: COVID-19, machine learning, imputation, XGBoost

INTRODUCTION

Since the COVID-19 outbreak began in 2019, the disease has spread all over the world and has caused serious public health concerns (Roncon et al. 2020). In addition, various variants of COVID-19 virus have appeared and the variants such as Delta and Omicron attract the special attention for public health (CDC 2020).
As it is highly infectious, many hospitals suffered from a surge of patients during the early stages, with some high-risk inpatients needing to be admitted to the ICU. However, not all such patients could be admitted to the ICU owing to the limited resources in ICUs. Owing to such scenarios, Aziz et al. (2020) recommended

ISSN: 1941-6679-On-line Copy

developing a mathematical model to predict the demand of ICUs for improved organizational management. Relating COVID-19, studies are conducted in variety of areas such as the understand of the sentiment of COVID-19 Tweets (Choudrie et al. 2021; Kaur et al. 2021), the detection of COVID-19 at early stage (Singh et al. 2021), COVID-19 vaccine allergic reaction (R. Schumaker et al. 2021), students' satisfaction of online class under COVID-19 pandemic (Bai et al. 2021), and potential COVID-19 drug interaction (R. P. Schumaker et al. 2021).

Most COVID-19 inpatients might not have detailed medical records at the moment of hospitalization, and there would also not be enough time to collect the patient's medical information. Upon the high demand of ICU resources due to a sudden surge of patients required to be admitted to ICU, the short-term prediction model for the patients who would be admitted to ICU would support the resource allocation and the seamless operation of ICUs. If we could develop a prediction method using limited sets of medical measurements, it would be beneficial to predict the necessity of COVID-19 inpatients' admission to ICU. The previous models predicting the future medical events including the COVID-19-related symptom changes and the general medical status changes have focused on relatively long-term (more than 15 days) predictions while the studies for short-term prediction focused on the prediction of the spread of COVID-19 and the overall demand change of ICUs. To the best of our knowledge, however, individual level models predicting the short-term ICU admission of a COVID-19 patient within several hours have not been developed. Therefore, to fill the gap, we propose a method predicting COVID-19 inpatients' admission to ICU within a short-term period.

In this study, we propose a method for predicting the COVID-19 inpatients' admission to ICU within a time frame of 12 hours. We use the dataset posted on Kaggle by a hospital in Brazil. The dataset includes 384 COVID-19 inpatients (324 patients' data were considered for this study), and each patient has five sets of measurements including patient demographic information, patient previous grouped diseases, and 216 derived features from 54 types of medical measures in five different time windows (from 0 to 2 hours, 2 to 4 hours, 4 to 6 hours, 6 to 12 hours, and above 12 hours from the time of hospital admission). We selected this dataset to establish a short-term prediction (within 12 hours) method for COVID-19 inpatients' ICU admission. For our method, we use a four-step process to train the classifier: the imputation of missing values using Bayesian Ridge Regression (BRR), the synthesis of training samples by the combination of two sets of measurements (the first time window with another one) of each patient, customized oversampling of training samples, and the classification using the XGBoost classifier. As the dataset has 55.4% missing values, we apply BRR to each feature of each patient to impute the missing values in the first step. In the next step, we synthesize the training samples by selecting the combinations of the first measurement and another measurement of the same patient and concatenating the

 ISSN: 1941-6679-On-line Copy

first measurement and the rates of changes from the first to another measurement into a sample. This step is conducted to increase the variability of training samples. In the third step, an oversampling is applied to mitigate the class imbalance problem on the synthesized training samples. Unlike typical oversampling, we selected at least one sample from each patient first and conducted random oversampling. As the last step, we train XGBoost by using the oversampled training set. In the experiments, we compare the Area Under the Curve-Receiver Operating Characteristics (AUC-ROC) and F-score of our methods with those of other methods using various pre-processing techniques and classifiers. Our method achieves the best performance among methods in the experiment. The result implies that the method proposed in this study (BRR imputation + synthesis of training samples using combinations of the first row and another row of each patient + customized oversampling + XGBoost) is viable for the prediction of the likelihood of COVID-19 inpatients' admission to the ICU and that the proposed imputation technique using BRR and the step-by-step approach to build methods are effective and applicable to other studies and industry practices. These are the contributions of the study and the practical usage of the method.

The rest of the paper is organized as follows: Section 2 presents the related works. In Section 3, the details of data used in the study are explained. The method is described in detail in Section 4. Section 5 includes the experiment setup, its result, and a discussion of the result. Section 6 presents the conclusion of this study.

RELATED WORKS

Studies for COVID-19

After the outbreak of COVID-19, the studies regarding COVID-19 have been conducted in various ways such as the prediction of COVID-19 cases, the diagnosis of COVID-19 using medical measures, the tracking of the changes of symptom and medical measures, and the model yielding risk score. In the perspective of time frame, the studies can be categorized into two groups, short-term (less than 15 days) and long-term (more than 15 days), and the various topics of studies are conducted in both groups.

In the short-term prediction, given that the trend of the number of COVID-19 cases is an important concern for the control of the disease, a large number of studies proposed prediction models for the spread of COVID-19 and the overall demand of ICUs (Bekker and Koole 2021; Berta et al. 2020; Bonnasse-Gahot et al. 2020; Català et al. 2020; Chin et al. 2020; El-Ghitany 2020; Farcomeni et al. 2021; Funk et al. 2020; Goic et al. 2021; Keeling et al. 2020; Manevski et al. 2020; Massonnaud et al. 2020; Petermann et al. 2021; Rahaman Khan and Hossain 2020;

Ricoca Peixoto et al. 2020; Weissman et al. 2020; C. Zhao et al. 2020; Zhao et al. 2021). However, these models predict the status changes of a population in particular regions (e.g., the short-term prediction of ICU demand in European countries). Another group of studies focused on medical measures. Computer tomography (CT) is adopted to predict the lung anomalies (Chassagnon et al. 2020), the CT score-based prognosis (Francone et al. 2020), and ventilation (Burdick et al. 2020) of COVID-19 patients in a short-term period and Solimando et al. (2021) and Vultaggio et al. (2020) tracked the short-term changes of Urea-to-Creatinine Ratios and IL-6 respectively.

In the long-term prediction, CT is adopted for the diagnosis of COVID-19, and the prediction models for the ICU demand, the risk of ICU admission, and the patient's symptom development were presented. Although there exist medical diagnostic methods such as performing polymerase chain reaction (PCR) test, ML-based diagnosis models using medical images such as Computed Tomography (CT) and X-ray have also been proposed (Ardakani et al. 2020; Barstugan et al. 2020; Elaziz et al. 2020; Ozturk et al. 2020). A prediction model using blood test results and the random forest algorithm was also proposed (Wu et al. 2020), which provided similar performance to the models using CT or symptom information for the diagnosis of COVID-19 patients.

Another concern with COVID-19 in the long-term prediction is the surge of patients. Many COVID-19 patients experience a deterioration in medical status after mild symptoms. This occurs in 26% to 32% COVID-19 patients admitted to the ICU (Huang et al. 2020; Wang et al. 2020). To predict the ICU demand for COVID-19 patients, several studies suggested ML models to stratify the COVID-19 patients. Gomes et al. (2020) proposed a model identifying the patients who will be developing severe symptoms and admitted to ICU. This model extracts features from X-ray images and analyzes the features using a decision tree to avoid overfitting. Roncon et al. (2020) also analyzed the risk of ICU admission and the mortality of diabetes patients. The result showed that patients having diabetes have a higher risk of ICU admission and death. In addition, patients having cardiovascular disease also have a higher risk of developing severe symptoms, requiring ICU admission and respiratory support treatment (He et al. 2020).

To predict the risk of ICU admission in the long-term, Z. Zhao et al. (2020) proposed the model yielding a risk score for the ICU admission and the mortality of COVID-19 patients. In the study, logistic regression is used to identify two outcome variables for ICU admission and mortality.

The risk score of the model yielded good accuracy with an AUC-ROC of 0.74 for the prediction of ICU admission and 0.83 for the prediction of mortality. This study was a retrospective study, with data being collected over approximately 1.5 months. Although their models do not directly predict ICU admission rates, some studies proposed models predicting COVID-19 patients' symptom development in the

 ISSN: 1941-6679-On-line Copy

long-term. Gao et al. (2020) provided a prediction model for physiological deterioration and death up to 20 days by adopting the ensemble of four different classifiers: logistic regression, support vector machines (SVM), gradient-boosted decision tree, and neural networks (NN). Di Castelnuovo et al. (2020) identified the characteristics of patients' in-hospital mortality due to COVID-19 by using random forest-based Cox survival analysis. Sun et al. (2020) identified 36 clinical indicators relevant to severe/critical symptoms and developed a prediction model using SVM. Assaf et al. (2020) proposed a model combining machine learning techniques and APACHE II risk prediction score for the prediction of the COVID-19 patients with non-severe, severe, and critical status. Machine learning techniques outperformed the baseline predictors based on a single variable or other risk predictions. Haase et al. (2020) identified the characteristics of COVID-19-related ICU patients such as median age, gender, chronic comorbidities, use of organ support, length of ICU stay, and vital signs.

Although the previous studies approached to COVID-19 in various perspectives such as the analysis CT and X-ray images, the tracking of the changes of medical measures and patients' symptom development, and the ICU demand prediction, the individual level model predicting the ICU admission of a COVID-19 inpatient was not proposed.

Availability of samples and missing data

One characteristic of medical data is the missing data created by irregular observation of patient status and varying measurement frequency (Lipton et al. 2016). In many datasets, it is known that missing values are correlated with the target variable (Rubin 1976). Therefore, a variety of missing-value imputation techniques such as forward/backward-filling, zero-imputation, regression imputation, multiple imputation (Rubin 1996), Multivariate Imputation by Chained Equation (MICE; Buuren & Groothuis-Oudshoorn (2010)), interpolation (Kreindler and Lumsden 2012), spline methods (De Boor and De Boor 1978, p. 27), k-NN algorithm (Song et al. 2008), and EM algorithm (García-Laencina et al. 2010) have been widely applied in practice. For the improved representation of EHRs, novel missing data techniques were proposed by Lipton et al. (2016). They proposed a new method by combining forward-filling and zero imputation. When imputing the values for the missing data, binary indicators were used to represent that the value is imputed by an imputation strategy. They also indicated that RNNs might recognize the fill-in values without indicators by learning the patterns of the fill-in values. GRU-D (Che et al. 2018) was also proposed for missing-value imputation. In addition to EHR-related variables, GRU-D receives masking and time intervals as input. This additional information enables the model to provide better prediction performance. BRR was also used to impute the missing values. In the approach in Mostafa et al. (2020), authors used values of complete features as inputs to fit a BRR model and used the fitted model to predict the missing values. BRR is formalized as follows (Mostafa et al. 2020):

$$v \sim N(\mu, \alpha), \tag{1}$$

where:

$$\mu = \beta X = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_n$$
$$\beta \sim N\,(0, \lambda^{-1} I_n)$$
$$\alpha \sim gamma\,(\alpha_1, \alpha_2)$$
$$\lambda \sim gamma\,(\lambda_1, \lambda_2)$$

Here, v is the target feature following Gaussian distribution, where variance is α and mean is $\mu = \beta X$, X is the independent features, β is the regression parameter of which independent Gaussian priors are variance $\lambda^{-1} I_p$ and mean 0, n is the number of independent features, λ and α are regularizing parameter of gamma distribution, and α_1, α_2, λ_1, and λ_2 are hyper-parameters of the gamma prior distributions.

XGBOOST

In recent years, diverse ML techniques such as linear and logistic regression, SVM, NN, and tree-based techniques have been applied in various classification and forecasting tasks. Among the techniques, Extreme Gradient Boosting (XGBoost) demonstrated successful performance in many competitions such as
Kaggle (Ogunleye and Wang 2020) and attracted considerable attention from researchers and practitioners (Gumus and Kiran 2017). XGBoost is based on Gradient Boosting (Friedman 2001) and uses tree learners as the base learner.
It was proposed by Chen & Guestrin (2016) to improve the performance of Gradient Boosting. For output prediction, XGBoost uses K additive functions as follows (Chen and Guestrin 2016; Nobre and Neves 2019):

$$\hat{y}_i = \emptyset(x_{\mathrm{i}}) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}, \tag{2}$$

where f is a function in the functional space $\mathcal{F}$, with $\mathcal{F} = \{f(x) = w_{q(x)}\}$ (q: $\mathbb{R}^m \longrightarrow T,\ w \in \mathbb{R}^T$) being the space of regression trees, where q is the structure of each tree that maps an example to the corresponding leaf index, T is the number of leaves in the tree, and w is the leaf weight.

Furthermore, it uses a loss function including a regularization term to avoid overfitting. The objective function $\mathcal{L}^{(t)}$and its regularization term $\Omega(f_t)$ are formalized as follows (Chen and Guestrin 2016; Nobre and Neves 2019):

$$\mathcal{L}^{(t)} = \sum_{i}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) \; + \; \Omega(f_t) \qquad (3)$$

In the objective function, $\hat{y}_i^{(t-1)}$ represents the prediction of the instance i at iteration t-1, $l(y_i, \hat{y}_i^{(t-1)})$ is the training loss function, and f_t is added to help the minimization of the objective (Nobre and Neves 2019). The second term Ω is for regularization. It helps avoid overfitting by penalizing the complexity of the model (Chen and Guestrin 2016). In addition to improved model performance and regularization, XGBoost provides a higher training speed by building trees parallely using multiple CPU cores even if it is based on a boosting method. In the perspective of dataset size, Rácz et al. (2021) conducted the performance test for five classifiers including XGBoost in three case studies. The classifiers employed for the study are XGBoost, naïve Bayes, SVM, multi-layer feed-forward of resilient backpropagation network, and probabilistic neural network (PNN). The sample sizes of the datasets were 100, 500, 1000, and all samples of the balanced dataset. In the 12 experiments, it showed the best performance in 11 experiments. This result implies that XGBoost yields high performance for the relatively small sample size datasets. Full details of XGBoost and its features are presented in Chen & Guestrin (2016).

Studies in various areas adopted XGBoost as the classifier. In the business domain, Nobre & Neves (2019) combined XGBoost with other pre-processing methods to forecast the trade in the finance market. Gumus & Kiran (2017) adopted XGBoost for crude oil price prediction. In the medical field, XGBoost was used for the classification of epilepsy patients (Torlay et al. 2017) and the diagnosis of Chronic Kidney Disease Diagnosis (Ogunleye and Wang 2020). Dhaliwal et al. (2018) used XGBoost for the Intrusion Detection System.

DATA DESCRIPTION

The dataset used in our experiment is a collection of anonymized COVID-19 inpatients' time series records of medical measurements from Hospital Sírio-Libanês, São Paulo and Brasilia. The data was anonymized by following international practices and posted at Kaggle (https://www.kaggle.com/S%C3%ADrio-Libanes/covid19).

The purpose of the creation of the dataset is to predict the likelihood of ICU admission of COVID-19 inpatients. The dataset was collected following the given scenario: once a COVID-19 patient is admitted to the hospital, medical staff conduct blood tests and measure vital signs until the patient is admitted to ICU or released from the hospital. The time windows used are: from 0 to 2 hours, 2 to 4 hours, 4 to 6 hours, 6 to 12 hours, and above 12 hours from the hospital admission. For each time window, they also record whether the patient is admitted to ICU during the time window.

Under such a scenario, 384 COVID-19 patients' records were provided in the dataset. Each row in the dataset includes a set of measurements of a patient during a particular time window. Each row has 228 variables, which are composed of three patient demographic information, nine patient previous grouped diseases, and 216 derived features from 54 types of medical measures such as albumin, oxygen saturation, heart rate, etc. by calculating the mean, median, max, min, difference of max and min, and relative difference (= difference/median).

As a patient's status is measured during five-time windows, each patient's measurement records appear in five rows. However, 55.4% values in the dataset are missing and the missing values appear irregularly in the dataset because blood tests and vital sign checks are conducted irregularly depending on the patient's status.

METHOD

The goal of the study is to propose a method predicting the likelihood of a COVID-19 inpatient's ICU admission based on the first two observations of the patient's status. The method is composed of three pre-processing steps and a classifier. As the dataset of the study has 55.4% missing values, the imputation of the missing values is conducted first by applying BRR to each feature of each patient separately. After the imputation, we synthesize the training samples to increase the variety of the training data. In this step, we select the combinations of the first row and another row of the same patient and concatenate the first row and the rates of changes of attributes between the two rows into a row. In the third step, we used a customized oversampling to mitigate the class imbalance between the two classes in the training dataset. After three pre-processing steps, we train the XGBoost classifier by using

 ISSN: 1941-6679-On-line Copy

the pre-processed training set. The high-level view of our method is provided in Figure 1.

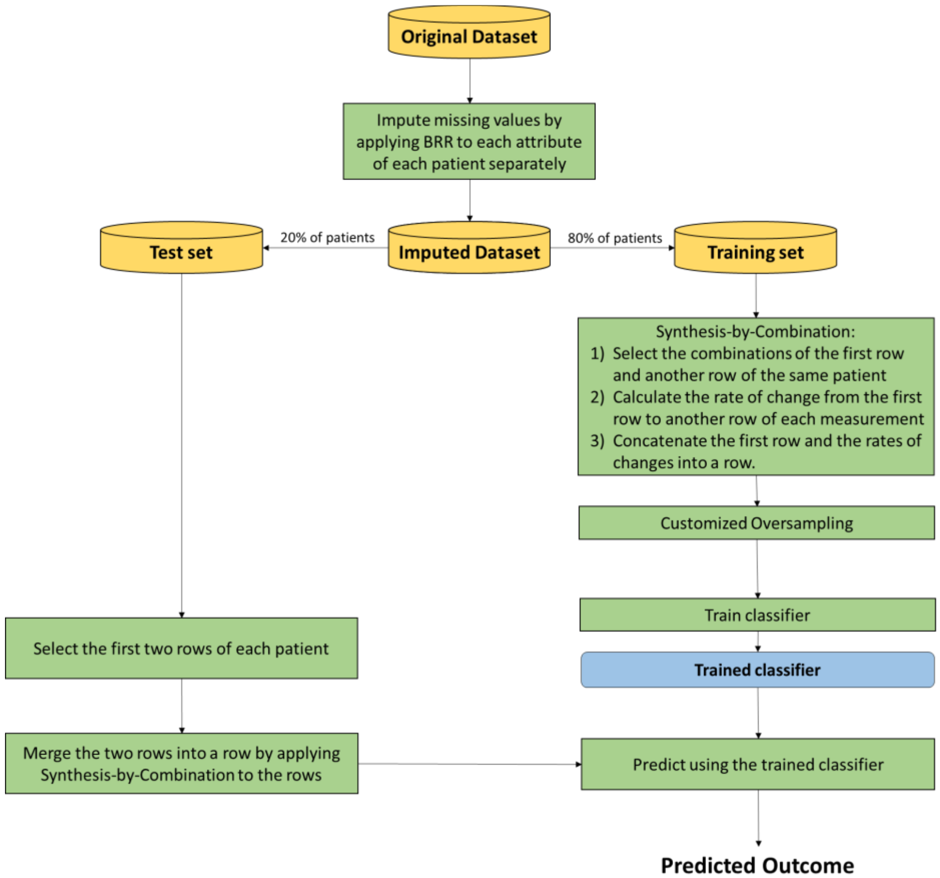

Fig. 1 Method of the study

Imputation of missing values

In the first step, we imputed the missing values in the dataset by applying BRR. BRR is selected based on the following scenario and assumption.

As per medical practices, if there are some changes in the patient's status, medical staff check or measure the patients' status and record them. Consequently, the measures such as temperature and heart rate of a patient appear at different intervals for different patients. Our assumption for the missing values is that if there is a missing value for a measurement in a time period then that value can be reasonably imputed based on the value from the preceding and following time periods.

For example, if a patient's body temperatures are 98°F, 101°F, and 104°F in the first-, third-, and fifth-time windows respectively, we might hypothesize that the temperatures would be approximately 99.5°F and 102.5°F in the second and fourth windows respectively with a high probability. In Mostafa et al. (2020), the BRR model is fitted by using multiple features as independent features to impute the missing values and once missing values of a feature is imputed, the imputed feature is used as an independent feature. However, in the application of this study, a BRR model is fitted by using only the values of each feature of each patient without modification of the BRR algorithm itself. In other words, we do not use imputed values in imputing other values. To the best of our knowledge, any previous studies have not used BRR in the manner used in this study.

The missing values are imputed by fitting the BRR models based on Gaussian probability distribution (Massaoudi et al. 2020) because it is adaptable to insufficient data. Suppose there are missing values of the feature *i* among *n* features of a patient p, $x_{p,i}^{miss}$ and complete values of the feature of the patient, $x_{p,i}^{comp}$.

The imputation of missing values of the feature *i* of patient *p* using BRR is formulated as follows:

$$x_{p,i}^{miss} \sim N(\mu, \alpha), \qquad (4)$$

where:

$$\mu = \beta X = \beta_0 + \beta_i x_{p,i}^{comp}$$
$$\beta \sim N\ (0, \lambda^{-1} I_n)$$
$$\alpha \sim gamma\ (\alpha_1, \alpha_2)$$
$$\lambda \sim gamma\ (\lambda_1, \lambda_2)$$

Here, β is the regression parameter of which independent Gaussian priors are mean 0 and variance $\lambda^{-1} I_n$, n is the number of independent features n=1, λ and α are regularizing parameters of gamma distribution set to $\alpha_1 = \alpha_2 = \lambda_1 = \lambda_2 = 10^{-6}$. The example of the imputation using BRR is presented in Tables 1 to 3.

Table 1 shows the status of the original data having missing values, in which missing values are represented as "Nan." A BRR model is fitted by using the values in ALBUMIN_MEDIAN column as a dependent variable and the values in TIME_WINDOW as an independent variable. After fitting the BRR model, the missing values on the ALBUMIN_MEDIAN column are imputed using the fitted BRR model.

Table 2 shows the status after imputing the missing values on ALBUMIN_MEDIAN column. In a similar way, the missing values in ALBUMIN_MEAN column can be imputed by fitting another BRR model. Table 3 depicts the status after imputing the missing values on ALBUMIN_MEAN column. The entire imputation algorithm of the missing values is provided in Figure 2.

Table 1. Before the imputation - Original Data

Patient ID	….	ALBUMIN_ MEDIAN	ALBUMIN_ MEAN	ALBUMIN_ MIN	…..	TIME_WINDOW
0		Nan	Nan	Nan		0
0		Nan	Nan	Nan		1
0		0.605263158	0.605263158	0.605263158		2
0		Nan	Nan	Nan		3
0		0	0	0		4

Table 2. After the imputation of ALBUMIN_MEDIAN column

Patient ID	….	ALBUMIN_ MEDIAN	ALBUMIN_ MEAN	ALBUMIN_ MIN	…..	TIME_WINDOW
0		1.210516	Nan	Nan		0
0		0.907888	Nan	Nan		1
0		0.605263	0.605263158	0.605263158		2
0		0.302632	Nan	Nan		3
0		0	0	0		4

Table 3. After the imputation of ALBUMIN_MEAN column

Patient ID	….	ALBUMIN_ MEDIAN	ALBUMIN_ MEAN	ALBUMIN_ MIN	…..	TIME_WINDOW
0		1.210516	1.210516	Nan		0
0		0.907888	0.907888	Nan		1
0		0.605263	0.605263	0.605263158		2
0		0.302632	0.302632	Nan		3
0		0	0	0		4

Input:

D: Dataset having complete and missing values

Output:

$D_{comp+\,imp}$: Dataset with all missing values imputed

Definitions:

P_{all}: List of all patients in the dataset

D_p: Rows of patient p

$\mathfrak{B}_{p,i}$: Bayesian Ridge Regression model for the feature i of patient p

$x_{p,i}^{comp}$: Complete values in the feature i of patient p

$x_{p,i}^{miss}$: Missing values in the feature i of patient p

ISSN: 1941-6679-On-line Copy

$x_{p,i}^{comp+imp}$: All data values of feature i of patient p in which all missing values are imputed

Begin
foreach patient $p \in P_{all}$ **do**
select D_p from D
foreach feature i in D_p **do**
fit $\mathfrak{B}_{p,i}$ using the values in $x_{p,i}^{comp}$
$x_{p,i}^{comp+imp}$ ← **impute** $x_{p,i}^{miss}$ **using** $\mathfrak{B}_{p,i}$
append $x_{p,i}^{comp+imp}$ **to** $D_{comp+imp}$
end foreach
end foreach
End

Fig. 2 Algorithm for the imputation of missing values

Synthesis of training samples using combinations (Synthesis-by-Combination)

Another issue in our dataset is the small number of samples. As the goal of our method is to predict the likelihood of COVID-19 inpatients' ICU admission based on the first two observations, we synthesize training samples by 1) resampling the combinations of the first row and another row of the same patient, 2) calculating the rate of change from the first row to another row of each feature, and 3) concatenating the first row and the rates of changes into a row. This not only increases the size of the data set but also provides more information to the classifier in terms of the rate of change of the different measurements.

This approach is based on the following assumption: some patients' status might be already severe at the first measurement, but others might be less severe than other patients even at the second measurement. In addition, the progression of the symptom can be different depending on the patient's physical condition. Therefore, if we could create combinations of two observations from each patient and concatenate them by calculating the rates of changes of measurements and train the classifier by using them, the classifier could be trained by a wider variety of cases. As this combination method requires at least two measurements for each patient, if only one measurement is available for a patient before ICU admission, that patient is not considered for the training set. The algorithm used for the creation of combinations of rows to create a training set is in Figure 3:

Input:
    $D_{comp+ imp}$: Dataset with all missing values imputed
Output:
    D_{synth}: Dataset synthesized using combination of two rows

Definitions:
    $P_{comp + imp}$: List of all patients in the dataset in which all missing values are imputed
    $D_{p,before}$: Rows of patient *p* created before the admission to ICU
    $D^{remain}_{p,before}$: The remaining rows after removing the first row from $D_{p,before}$
    X_{first}: The first row of a patient
    X_{second}: The row considered the later measurement row of a patient
    $\Delta X_{first, second}$: *The rates of changes of measures between* X_{first} *and* X_{second}

Begin
    foreach patient $p \in P_{comp+ imp}$ **do**
        select $D_{p, before}$ from $D_{comp+ imp}$
        if the number of rows in $D_{p, before}$ >= 2 **do**
            X_{first} ← **select** the first measurement row in $D_{p, before}$
            $D^{remain}_{p,before}$ ← **remove** the first measurement row from $D_{p, before}$
            foreach row X_{second} in $D^{remain}_{p,before}$ **do**
                $\Delta X_{first, second}$ ← calculate the rates of changes of measurements between X_{first} and X_{second}

                append concatenate (X_{first}, $\Delta X_{first, second}$) **to** D_{synth}

            end foreach
        end if
    end foreach

End

Fig. 3 **Algorithm for the synthesis of training samples**

Customized Oversampling

Resampling techniques such as oversampling and undersampling are widely used to mitigate the class imbalance problem. In this study, we use oversampling to address the class imbalance problem in the training set synthesized in the previous step. While oversampling generally uses random sampling, we modified the oversampling method so that at least one measurement from each patient is included in the training set.

XGBoost binary classifier

We select one data point for each patient in the minor class first and then randomly resample more samples from the minor class so that the number of data points in both classes are the same.

We employ the XGBoost classifier for learning the prediction model because of two reasons; better performance of XGBoost with smaller dataset (Rácz et al. 2021) and the prevention of overfitting by regularization (Chen and Guestrin 2016). The output of the classifier is the prediction of a COVID-19 patient's ICU admission and is within the range [0,1]. Scikit-learn machine learning library is used to implement the method. Although it is possible to find the optimal hyper-parameter by grid search, we use the default hyper-parameter values in the library so that we can measure the effect of the classifier and compare the performance of XGBoost with those of other classifiers. Important hyper-parameters used are: learning rate=0.3, subsample=1, maximum depth of tree=6, sampling method=uniform, lambda=1 (L2 regularization term on weights), and alpha=0 (L1 regularization term on weights).

EXPERIMENT AND DISCUSSION

Experimental Setup

We compared the performance of our method with five other methods using different classifiers and different combinations of pre-processing techniques. We employed six classifiers first and selected the best performing technique in each pre-processing step of each classifier by conducting the experiments separately for each classifier and each pre-processing step. After selection of pre-processing steps of five competing classifiers, our method is compared with these five methods.

The five classifiers that are compared to XGBoost are decision tree, SVM, Multi-layer Perceptron (MLP), AdaBoost, and random forest. They are selected based on their popularity, performance, and type of algorithm such as a basic classifier, function-based classifier, and ensemble classifier. The various imputation techniques we compared are zero filling, forward filling + zero filling (Lipton et al. 2016), backward filling + forward filling, forward filling + backward filling, and BRR. All imputation techniques are selected based on the assumption that missing values would be related to previous or subsequent values. However, the strategies for filling the remaining missing values after the application of the first technique are different.

For example, forward filling + zero filling technique imputes the missing values by forward filling first and applies the zero filling next if there still exist missing

values. For combination techniques, we compared three variations: No combination, combination of first and other measurements of each patient, and all possible combinations of all available measurements of each patient. No combination option does not synthesize any training samples. Combination of first and other measurements of each patient option is to use the combinations of the first measurement with another measurement of each patient to synthesize training samples. The details of this option are provided in section 4.2. All possible combinations of each patient option synthesizes the training samples by randomly picking any two measurements of that patient and combining them. In the resampling step, we compared three options: No resampling, Oversampling, and Undersampling. No resampling option does not apply resampling to the training set while Oversampling and Undersampling apply the samplings to the synthesized training set. As described in section 4.3, Oversampling selects at least one the minor class samples from each patient first and resample more minor samples so that both classes have the same number of samples while the Undersampling selects at least one major class samples from each patient first and resample more major samples.
To choose the best pre-processing techniques for each method, we tested each pre-processing step of each classifier ten times and compared its performance as measured by AUC-ROC and F-score. Since the goal of our method is to predict the likelihood of ICU admission for each patient by using the first two sets of a patient's measurements, we selected data for only those patients having at least two sets of measurements (two rows) measured before admission to the ICU. The number of such patients was 324. In each experiment, 20% of randomly selected patients are used as the test set.
All the algorithms and classifiers and the imputation, the training sample synthesis, and resampling techniques in all experiments are implemented using Scikit-learn machine learning library 1.0 and Python 3.8.

EXPERIMENTAL RESULTS

Selection of imputation technique

We compared five imputation techniques and trained them with six classifiers separately. The results of the experiment are presented in Tables 4 to 9. Note that all results presented below are average of 10 runs. In Table 4, the results of the experiments with five imputation techniques and decision tree are listed.
When combined with BRR, the decision tree provided the best AUC-ROC (0.681) and the best F-score (0.699). From this result, we selected the BRR as the imputation technique that will be used for the decision tree.

Table 4. Performance of imputation techniques with decision tree

Classifier	Imputation Technique	Total Accuracy	Accuracy for the class “Admit”	Accuracy for the class “Not Admit”	AUC-ROC	Precision	Recall	F-score
Decision tree	zero filling	0.597	0.547	0.640	0.629	0.550	0.547	0.536
	forward filling + zero filling	0.603	0.448	0.712	0.592	0.526	0.448	0.476
	backward filling + forward filling	0.668	0.479	0.798	0.628	0.620	0.479	0.535
	forward filling + backward filling	0.663	0.584	0.723	0.634	0.599	0.584	0.584
	BRR	0.745	0.629	0.826	**0.681**	0.720	0.629	**0.669**

Table 5. Performance of imputation techniques with random forest

Classifier	Imputation Technique	Total Accuracy	Accuracy for the class “Admit”	Accuracy for the class “Not Admit”	AUC-ROC	Precision	Recall	F-score
Random forest	zero filling	0.597	0.245	0.877	0.635	0.622	0.245	0.341
	forward filling + zero filling	0.591	0.246	0.853	0.584	0.588	0.246	0.328
	backward filling + forward filling	0.614	0.281	0.892	0.704	0.698	0.281	0.392
	forward filling + backward filling	0.655	0.314	0.901	0.700	0.689	0.314	0.424
	BRR	0.755	0.538	0.909	**0.816**	0.802	0.538	**0.641**

In Table 5, the results of the experiments with five imputation techniques and random forest are presented, and the BRR yielded the best AUC-ROC (0.816) and F-score (0.641) for the random forest classifier. Based on the result, we selected the BRR as the imputation technique that will be used for the random forest.

Similar to previous results, the BRR also yielded the best AUC-ROCs and F-scores when combined with MLP (AUC-ROC = 0.877 and F-score = 0.754), SVM (AUC-ROC = 0.871 and F-score = 0.749), AdaBoost (AUC-ROC = 0.874 and F-score = 0.763) respectively as in Table 6 to 8. According to the results, we selected the BRR as the imputation techniques for MLP, SVM, and AdaBoost

 ISSN: 1941-6679-On-line Copy

Table 6. Performance of imputation techniques with MLP

Classifier	Imputation Technique	Total Accuracy	Accuracy for the class "Admit"	Accuracy for the class "Not Admit"	AUC-ROC	Precision	Recall	F-score
MLP	zero filling	0.642	0.484	0.763	0.695	0.607	0.484	0.532
	forward filling + zero filling	0.635	0.506	0.727	0.708	0.578	0.506	0.534
	backward filling + forward filling	0.672	0.524	0.777	0.730	0.636	0.524	0.570
	forward filling + backward filling	0.665	0.510	0.779	0.729	0.623	0.510	0.551
	BRR	0.799	0.719	0.864	**0.877**	0.810	0.719	**0.754**

Table 7. Performance of imputation techniques with SVM

Classifier	Imputation Technique	Total Accuracy	Accuracy for the class "Admit"	Accuracy for the class "Not Admit"	AUC-ROC	Precision	Recall	F-score
SVM	zero filling	0.649	0.425	0.810	0.706	0.604	0.425	0.490
	forward filling + zero filling	0.659	0.412	0.840	0.736	0.668	0.412	0.498
	backward filling + forward filling	0.714	0.563	0.815	0.774	0.668	0.563	0.605
	forward filling + backward filling	0.715	0.556	0.839	0.777	0.724	0.556	0.621
	BRR	0.805	0.707	0.877	**0.871**	0.803	0.707	**0.749**

ISSN: 1941-6679-On-line Copy

Table 8. Performance of imputation techniques with AdaBoost

Classifier	Imputation Technique	Total Accuracy	Accuracy for the class "Admit"	Accuracy for the class "Not Admit"	AUC-ROC	Precision	Recall	F-score
AdaBoost	zero filling	0.665	0.539	0.753	0.663	0.613	0.539	0.568
	forward filling + zero filling	0.657	0.558	0.726	0.690	0.589	0.558	0.570
	backward filling + forward filling	0.662	0.536	0.768	0.717	0.659	0.536	0.585
	forward filling + backward filling	0.682	0.549	0.770	0.727	0.612	0.549	0.574
	BRR	0.806	0.761	0.839	**0.874**	0.773	0.761	**0.763**

XGBoost, the classifier of the method of the study, also provided the best performance when combined with BRR (AUC-ROC = 0.905 and F-score = 0.816) as shown in Table 9, so BRR is selected as the imputation technique for XGBoost. In the experiments for the imputation technique selection, the combination of BRR and XGBoost showed the AUC-ROC over 0.9 and F-score over 0.8 while BRR and other classifiers did not achieve 0.9 of AUC-ROC and 0.8 of F-score.

Table 9. Performance of imputation techniques with XGBoost

Classifier	Imputation Technique	Total Accuracy	Accuracy for the class "Admit"	Accuracy for the class "Not Admit"	AUC-ROC	Precision	Recall	F-score
XGBoost	zero filling	0.674	0.518	0.794	0.733	0.650	0.518	0.573
	forward filling + zero filling	0.657	0.510	0.762	0.707	0.609	0.510	0.552
	backward filling + forward filling	0.705	0.578	0.802	0.782	0.675	0.578	0.616
	forward filling + backward filling	0.666	0.526	0.784	0.755	0.668	0.526	0.583
	BRR	0.846	0.805	0.878	**0.905**	0.831	0.805	**0.816**

As shown in the tables, all classifiers yield the best AUC-ROC and F-score at the same time when combined with BRR. Based on the results, BRR is selected as the imputation technique for all classifiers in the experiment.

 ISSN: 1941-6679-On-line Copy

Selection of the training sample synthesis technique

After the selection of imputation techniques, we conducted the experiments for the selection of the training sample synthesis technique. As the BRR is selected for the imputation technique for all classifiers, we imputed the missing values by BRR for all classifiers first and applied three options for the synthesis of training samples to each classifier separately. The results are presented in Tables 10 to 15. In the experiments for BRR + three synthesis techniques + decision tree in Table 10, the method provided the best AUC-ROC (0.749) and the best F-score (0.705) when combined with Combination of first and other measurements of each patient technique. Based on the result, we adopted the Combination of first and other rows of each patient technique for the pre-preprocessing step for the decision tree.

In the experiments for random forest in Table 11, we selected the No combination of rows option as the training samples synthesis technique. As the best AUC-ROC (0.852) is provided when All possible combinations of all available rows of each patient option is applied and the best F-score (0.641) was achieved when No combination of rows option is selected, we considered the gap between the measures. While AUC-ROC of No combination of rows option was the lowest among three AUC-ROCs and the gap between the best one was 0.036, the F-score of No combination of rows option was the best among three options and the gap between the that of All possible combination of all available rows was 0.237. The gaps of two measures imply that No combination of rows option provides more stable performance than other options. Based on the analysis, we selected the No combination of rows option for the random forest.

In the experiments for MLP and SVM, the No combination of rows option yielded the best performance at the same time in AUC-ROC and F-score. The AUC-ROC for MLP was 0.877 and F-score was 0.754 and the AUC-ROC for SVM was 0.871 and the F-score was 0.749. For both classifiers, the No combination of rows option is selected as the training samples synthesis technique. The results are presented in Tables 12 and 13. In the experiments for AdaBoost, the Combination of first and other rows of each patient technique is selected based on the best AUC-ROC (0.888) and the best F-score (0.772). The result is presented in Table 14.

As the synthesis technique for XGBoost, the Combination of first and other rows option was selected according to the result shown in Table 15. The best AUC-ROC (0.919) was provided when the Combination of first and other rows of each patient option was applied, but the best F-score (0.816) was achieved by the No combination of rows option. When considering the gap between the best and the second AUC-ROC, it was 0.014, but the gap between the best F-score from the No combination of rows option and the one from the Combination of first and other rows of each patient option was 0.03. Although both gaps are not big, as the gap of AUC-ROC was greater than that of F-score, we selected the Combination of first and other rows of each patient option as the training sample synthesis.

Table 10. Performance of row synthesis techniques with decision tree

Classifier	Imputation Technique	Combination of Rows	Total Accuracy	Accuracy for the class "Admit"	Accuracy for the class "Not Admit"	AUC-ROC	Precision	Recall	F-score
Decision tree	**BRR**	No combination of rows	0.745	0.629	0.826	0.681	0.720	0.629	0.669
		All possible combinations of all available rows	0.760	0.545	0.897	0.707	0.781	0.545	0.632
		Combination of first and other rows	0.789	0.674	0.867	**0.749**	0.753	0.674	**0.705**

Table 11. Performance of row synthesis techniques with random forest

Classifier	Imputation Technique	Combination of Rows	Total Accuracy	Accuracy for the class "Admit"	Accuracy for the class "Not Admit"	AUC-ROC	Precision	Recall	F-score
Random forest	**BRR**	**No combination of rows**	0.755	0.538	0.909	0.816	0.802	0.538	**0.641**
		All possible combinations of all available rows	0.683	0.263	0.992	**0.852**	0.959	0.263	0.404
		Combination of first and other rows	0.723	0.343	0.987	0.835	0.948	0.343	0.499

Table 12. Performance of row synthesis techniques with MLP

Classifier	Imputation Technique	Combination of Rows	Total Accuracy	Accuracy for the class "Admit"	Accuracy for the class "Not Admit"	AUC-ROC	Precision	Recall	F-score
MLP	**BRR**	**No combination of rows**	0.799	0.719	0.864	**0.877**	0.810	0.719	**0.754**
		All possible combinations of all available rows	0.772	0.608	0.893	0.848	0.825	0.608	0.694
		Combination of first and other rows	0.783	0.644	0.895	0.863	0.827	0.644	0.719

ISSN: 1941-6679-On-line Copy

Table 13. Performance of row synthesis techniques with SVM

Classifier	Imputation Technique	Combination of Rows	Total Accuracy	Accuracy for the class "Admit"	Accuracy for the class "Not Admit"	AUC-ROC	Precision	Recall	F-score
SVM	**BRR**	**No combination of rows**	0.805	0.707	0.877	**0.871**	0.803	0.707	**0.749**
		All possible combinations of all available rows	0.789	0.637	0.895	0.854	0.809	0.637	0.710
		Combination of first and other rows	0.769	0.612	0.882	0.829	0.790	0.612	0.687

Table 14. Performance of row synthesis techniques with AdaBoost

Classifier	Imputation Technique	Combination of Rows	Total Accuracy	Accuracy for the class "Admit"	Accuracy for the class "Not Admit"	AUC-ROC	Precision	Recall	F-score
AdaBoost	**BRR**	No combination of rows	0.806	0.761	0.839	0.874	0.773	0.761	0.763
		All possible combinations of all available rows	0.791	0.653	0.891	0.851	0.805	0.653	0.718
		Combination of first and other rows	0.814	0.756	0.861	**0.888**	0.802	0.756	**0.772**

Table 15. Performance of row synthesis techniques with XGBoost

Classifier	Imputation Technique	Combination of Rows	Total Accuracy	Accuracy for the class "Admit"	Accuracy for the class "Not Admit"	AUC-ROC	Precision	Recall	F-score
XGBoost	**BRR**	No combination of rows	0.846	0.805	0.878	0.905	0.831	0.805	**0.816**
		All possible combinations of all available rows	0.832	0.674	0.947	0.907	0.903	0.674	0.770
		Combination of first and other rows	0.855	0.757	0.930	**0.919**	0.886	0.757	0.813

ISSN: 1941-6679-On-line Copy

Selection of resampling technique

For the last experiment, we compared different resampling techniques for each classifier. The pre-processing techniques selected in the previous sections are applied to each classifier first and three resampling techniques are applied as the last option for each classifier. The results are presented in Tables 16 to 21. In Table 16, the result of the experiment for decision tree was presented. The best AUC-ROC (0.749) appeared when No resampling option was selected while the best F-score (0.731) was achieved when Undersampling is applied. From the result, we selected Undersampling as the resampling technique because the gap of F-score (0.026) between No resampling and Undersampling options is greater than that of AUC-ROC (0.08) between the same options.
In the experiments for random forest presented in Table 17, the best AUC-ROC (0.835) and F-score (0.737) achieved when Oversampling is applied as the resampling technique. Therefore, the Oversampling is selected for the resampling technique for the random forest. For MLP, the Undersampling is selected as the resampling technique based on the best AUC-ROC (0.878) and the best F-score (0.759), and, for SVM, the No resampling option is selected based on the best AUC-ROC (0.871) and the best F-score (0.749). The results for three classifiers are presented in Tables 17-19.
The result of the experiment for AdaBoost in Table 20 showed the best AUC-ROC (0.888) and the best F-score (0.772) with No resampling option. Therefore, we selected the No resampling option for AdaBoost. Also, in the experiment for XGBoost shown in Table 21, Oversampling is selected as the resampling technique as it yielded the best AUC-ROC (0.932) and the best (0.840).

Comparison of Methods

As shown in the experiments in the sections 5.2.1 to 5.2.3, we selected best combination of the pre-processing steps for each of the six classifiers. In Table 22, the improvement of AUC-ROCs and F-scores of the selected methods in each step are summarized. The AUC-ROCs and F-scores improve or stay the same with each additional pre-processing step for almost all of the methods.

Table 23 summarizes the performance of six final methods discussed before. The best AUC-ROC (0.932) and the best F-score (0.840) are achieved by the BRR + Combination of first and other rows + Oversampling + XGBoost. The performance improvement in AUC-ROC and F-score of the best method compared to other methods is also shown to be statistically significant at 0.05 level of significance by the paired t-test. The table shows the corresponding p-values.

Table 16. Performance of row resampling techniques with decision tree

Classifier	Imputation Technique	Combination of Rows	Resampling Technique	Total Accuracy	Accuracy for the class "Admit"	Accuracy for the class "Not Admit"	AUC-ROC	Precision	Recall	F-score
Decision tree	**BRR**	**Combination of first and other rows**	No resampling	0.789	0.674	0.867	**0.749**	0.753	0.674	0.705
			Under sampling	0.780	0.733	0.816	0.741	0.732	0.733	**0.731**
			Over sampling	0.771	0.718	0.803	0.731	0.701	0.718	0.706

Table 17. Performance of row resampling techniques with random forest

Classifier	Imputation Technique	Combination of Rows	Resampling Technique	Total Accuracy	Accuracy for the class "Admit"	Accuracy for the class "Not Admit"	AUC-ROC	Precision	Recall	F-score
Random forest	**BRR**	**No combination of rows**	No resampling	0.755	0.538	0.909	0.816	0.802	0.538	0.641
			Under sampling	0.735	0.520	0.908	0.833	0.826	0.520	0.633
			Over sampling	0.792	0.690	0.867	**0.835**	0.798	0.690	**0.737**

Table 18. Performance of row resampling techniques with MLP

Classifier	Imputation Technique	Combination of Rows	Resampling Technique	Total Accuracy	Accuracy for the class "Admit"	Accuracy for the class "Not Admit"	AUC-ROC	Precision	Recall	F-score
MLP	**BRR**	**No combination of rows**	No resampling	0.799	0.719	0.864	0.877	0.810	0.719	0.754
			Under sampling	0.812	0.722	0.880	**0.878**	0.811	0.722	**0.759**
			Over sampling	0.789	0.674	0.870	0.859	0.786	0.674	0.724

ISSN: 1941-6679-On-line Copy

Table 19. Performance of row resampling techniques with SVM

Classifier	Imputation Technique	Combination of Rows	Resampling Technique	Total Accuracy	Accuracy for the class "Admit"	Accuracy for the class "Not Admit"	AUC-ROC	Precision	Recall	F-score
SVM	**BRR**	**No combination of rows**	**No resampling**	0.805	0.707	0.877	**0.871**	0.803	0.707	**0.749**
			Under sampling	0.788	0.645	0.898	0.855	0.814	0.645	0.710
			Over sampling	0.797	0.765	0.817	0.861	0.735	0.765	0.746

Table 20. Performance of row resampling techniques with AdaBoost

Classifier	Imputation Technique	Combination of Rows	Resampling Technique	Total Accuracy	Accuracy for the class "Admit"	Accuracy for the class "Not Admit"	AUC-ROC	Precision	Recall	F-score
AdaBoost	**BRR**	**Combination of first and other rows**	**No resampling**	0.814	0.756	0.861	**0.888**	0.802	0.756	**0.772**
			Under sampling	0.814	0.764	0.849	0.879	0.783	0.764	0.771
			Over sampling	0.791	0.739	0.833	0.850	0.765	0.739	0.748

Table 21. Performance of row resampling techniques with XGBoost

Classifier	Imputation Technique	Combination of Rows	Resampling Technique	Total Accuracy	Accuracy for the class "Admit"	Accuracy for the class "Not Admit"	AUC-ROC	Precision	Recall	F-score
XGBoost	**BRR**	**Combination of first and other rows**	No resampling	0.855	0.757	0.930	0.919	0.886	0.757	0.813
			Under sampling	0.840	0.770	0.896	0.901	0.838	0.770	0.799
			Over sampling	0.869	0.826	0.904	**0.932**	0.861	0.826	**0.840**

ISSN: 1941-6679-On-line Copy

Table 22. Performance change by each classifier and each step

	AUC-ROC			F-score		
Classifier	**Imputation only**	**Imputation + Combination**	**Imputation + Combination + Resampling**	**Imputation only**	**Imputation + Combination**	**Imputation + Combination + Resampling**
Decision tree	0.681	0.749	0.749	0.669	0.705	0.731
Random forest	0.816	0.816	0.835	0.641	0.641	0.737
MLP	0.877	0.877	0.878	0.754	0.754	0.759
SVM	0.871	0.871	0.871	0.749	0.749	0.749
AdaBoost	0.874	0.888	0.888	0.763	0.772	0.772
XGBoost	0.905	0.919	**0.932**	0.816	0.813	**0.840**

Table 23. Performance of all methods

Classifier	Imputation Technique	Combination of Rows	Resampling Technique	Total Accuracy	Accuracy for the class "Admit"	Accuracy for the class "Not Admit"	AUC-ROC	Precision	Recall	F-score
Decision Tree	**BRR**	**Combination of first and other rows**	**Under sampling**	0.780	0.733	0.816	0.741 (p=.000)	0.732	0.733	0.731 (p=.000)
Random forest	**BRR**	**No combination of rows**	**Over sampling**	0.792	0.690	0.867	0.835 (p=.000)	0.798	0.690	0.737 (p=.000)
MLP	**BRR**	**No combination of rows**	**Under sampling**	0.812	0.722	0.880	0.878 (p=.001)	0.811	0.722	0.759 (p=.006)
SVM	**BRR**	**No combination of rows**	**No resampling**	0.805	0.707	0.877	0.871 (p=.000)	0.803	0.707	0.749 (p=.001)
AdaBoost	**BRR**	**Combination of first and other rows**	**No resampling**	0.814	0.756	0.861	0.888 (p=.016)	0.802	0.756	0.772 (p=.039)
XGBoost	**BRR**	**Combination of first and other rows**	**Over sampling**	0.869	0.826	0.904	**0.932**	0.861	0.826	**0.840**

ISSN: 1941-6679-On-line Copy

Explainibility of the model

In the applications of ML models, one criteria that is becoming important is the explainibility of the model; so called Explainable AI. As XGBoost is a decision-tree based ensemble algorithm, the advantage of using it is that the interpretation of its models is relatively easy compared to other algorithms such as MLP and Convolutional Neural Network (CNN) while also providing a good performance. For the understanding of our model, we present a decision tree and an importance plot from the trained XGBoost model of our method (BRR + Combination of first and other rows + Oversampling + XGBoost) after one run. In Figure 4, we show the first order tree among the trees created by XGBoost in an additive way. Each tree has a depth of 6 (defined in the parameter setting), but for the sake of simplicity we only show the tree till depth 3.

We can see that the trained XGBoost model starts its decision tree from "RATE_RESPIRATORY_RATE_MAX" attribute and extends the decision criteria to "RATE_LACTATE_MEDIAN" and "BLOODPRESSURE_DIASTOLIC_MEDIAN" attributes. Note the features names starting with "RATE_" are the attributes we created by calculating the changing rate of each feature from the first measurement.

While the trees provide the advantages to observe the detailed decision criteria, it is not straightforward to recognize which attributes are relatively more effective for the decision. Therefore, we presented an importance plot in Figure 5, which is plotted using the fitted XGBoost model of our method after a run of experiment based on the information gain of each attribute. In the plot, the AGE_PERCENTILE attribute is considered most important by showing the largest gain among the attributes. This is aligned with the fact that the elderly showed higher mortality rate than younger people (Di Castelnuovo et al. 2020). The other attributes in order of importance were, BILLIRUBIN_MEDIAN, DIMER_MEDIAN, RATE_RASPITORY_RATE_DIFF, and AGE_ABOVE.

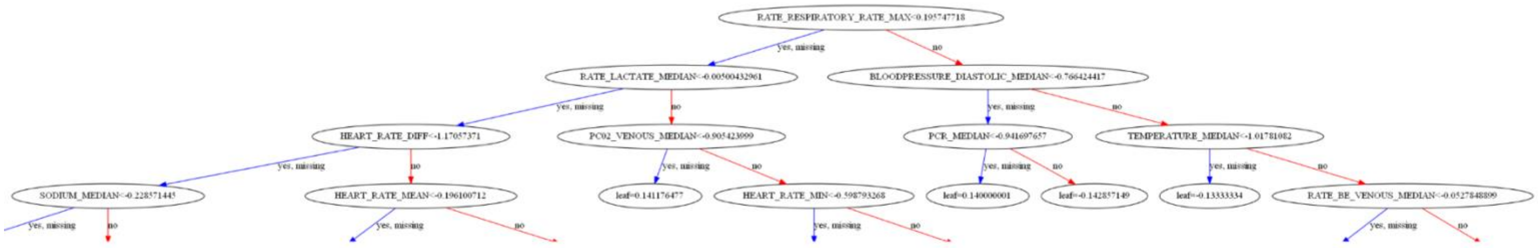

Fig. 4 The first order tree created from the fitted XGBoost model in the method of this study

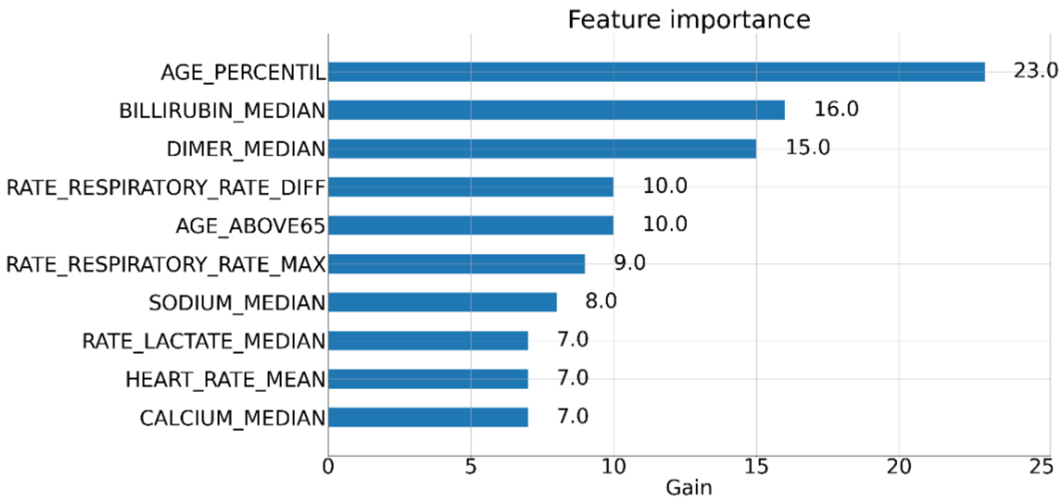

Fig. 5 An importance plot created from the fitted XGBoost model in the method of this study

DISCUSSION

Using incremental comparison, we found the combination of pre-processing steps and classifier of BRR + Synthesis-by-Combination + Oversampling + XGBoost to provide the best performance (AUC-ROC = 0.932 and F-score = 0.840) among the methods in the experiments. The results of the step-by-step experiments implies that a proper combination of techniques and a classifier could improve the performance of a method even if the same set of techniques are used. In our method, while all steps contribute to the improvement of overall performance of the method, BRR and XGBoost play a critical role for the performance improvement. As observed in the Tables 10 to 21, all methods combined with BRR and XGBoost achieved the AUC-ROC greater than 0.9 while other classifiers did not achieve that level of performance even when using BRR imputation.

In addition to the employment of BRR and XGBoost, the development and the application of Synthesis-by-Combination technique also contributed to the performance improvement. The list of 10 most important attributes in Figure 5 includes three attributes created by Synthesis-by-Combination (the attributes starting with “RATE_”). It means that the changing rates of certain attributes are important criteria deciding a patient’s admission to ICU and that information is provided by Synthesis-by-Combination. We hypothesize the reasons for the performance improvement in the experiment to be: 1) our assumption for the missing values was correct and the imputation technique proposed in this study implements the assumption correctly by employing BRR,2) XGBoost provides a good performance for the smaller datasets as indicated in Rácz et al. (2021) and

ISSN: 1941-6679-On-line Copy

effectively prevents the overfitting that could be caused by smaller datasets by adopting the regularization (Chen and Guestrin 2016), and 3) the Synthesis-by-Combination provides additional piece of information for the decision making of classifier.

In addition to the implications from the entire method, there are a couple of additional implications of the results. First, the proposed imputation technique using BRR is effective even when used with other classifiers. In the results for imputation techniques presented in Tables 4 to 9, the BRR imputation achieved the best AUC-ROCs and F-scores for all classifiers. The results imply that the BRR imputation technique could be applied to other datasets having similar structure. Secondly, the approach made in this study to build the best method for each classifier can be applied to other domains. Although the combination of BRR + Synthesis-by-Combination + Oversampling + XGBoost achieved the best performance, the best performing combinations of pre-processing steps for each classifier are different. It means that a particular set of pre-processing steps does not always yield the best result with all classifiers. The result in Table 22 implies that our approach to build methods is an effective one to find the best working method for each classifier, and it could be applied in other applications.

CONCLUSION

When a highly infective and dangerous disease with unknown characteristics starts spreading, one of the critical concerns related to hospital operation is the organizational management of limited ICU resources (Aziz et al. 2020). During the COVID-19 outbreak, hospitals suffered from the surge of patients requiring ICU admission. Aziz et al. (2020) suggested using mathematical modeling to predict the surge of patients for better resource allocation. To tackle the issue, we develop a method predicting the COVID-19 inpatients' likelihood of ICU admission based on two sets of patient measurements. We use the dataset collected during the early stage of the outbreak of COVID-19 from a hospital in Brazil. As the dataset has 55.4% missing values and only 324 available patients, we developed a method involving three pre-processing steps before a classifier is used. In the first step, the missing values are imputed by applying an imputation technique using BRR to each attribute of each patient separately. In the second step, we increase the variability of the training set by synthesizing new samples using the combinations of the first and later measurements of each patient. In the third step, we create a customized undersampling, which samples at least one sample of each patient first and conducts the random sampling to balance the classes, to mitigate the class imbalance problem of the synthesized training set. As the final step of the method, XGBoost classifier is used to predict the likelihood of ICU admission for each patient. To select the best combination of methods, we conducted step-by-step experiments and built five competing methods using different sets of pre-processing steps and classifiers. We compared the AUC-ROC and F-score of our method to those of others. Among the various methods considered in this study, our method provides the highest AUC-ROC (0.932) and F-score (0.840). In addition, our method showed that the AGE_PERCENTILE is the most important attribute affecting patients' admission to ICU. Based on the result of the experiment, we may summarize the contribution of the study as follows: 1) combination of BRR imputation, Synthesis-by-Combination, Oversampling, and XGBoost provides the best performance for the prediction of the likelihood of ICU admission of COVID-19 inpatients, 2) the proposed imputation technique using BRR in this study is an effective one to impute the missing values, and 3) the proposed step-by-step approach could be adopted for the method selection for other domains or diseases.

Although our method demonstrated good performance, there is room for the improvements in the future studies. Medical data has several missing values owing to the characteristics of medical record creation procedures. The imputation techniques for medical data considering the characteristics of medical procedure should be developed or further improved because, as observed in our study, better imputation techniques for medical data could improve the model performance.

Also, more sophisticated and generalized model is required for the detailed resource planning of ICUs. One of the reasons we picked this dataset was because of the challenging problem it posed. Not only was the dataset relatively small but it had missing values in the sense that different patients had their vital readings taken in different time periods without any uniformity. We came up with the unique method of Synthesis-by-Combination described in section 4.2 that not only overcame this problem by creating a uniform data set but also increased the number of data points in the data sets.

Given the small size of the data set there is also the question of generalizability of the model that is learned. Given that different parts of the world were dealing with slightly different strains of Covid-19 it is possible that the models learned in one part of the world may not apply to other regions any way. However, we agree that if we could use a larger dataset collected from more patients, it is possible that the generalizability of the model will be improved and the model will be able to predict the exact time range of ICU admission. As the result, the efficiency of ICU operation will also be enhanced.

REFERENCES

Ardakani, A. A., Kanafi, A. R., Acharya, U. R., Khadem, N., and Mohammadi, A. 2020. "Application of Deep Learning Technique to Manage COVID-19 in Routine Clinical Practice Using CT Images: Results of 10 Convolutional Neural Networks,"

ISSN: 1941-6679-On-line Copy

Computers in Biology and Medicine (121), p. 103795. (https://doi.org/10.1016/j.compbiomed.2020.103795).

Assaf, D., Gutman, Y., Neuman, Y., Segal, G., Amit, S., Gefen-Halevi, S., Shilo, N., Epstein, A., Mor-Cohen, R., Biber, A., Rahav, G., Levy, I., and Tirosh, A. 2020. "Utilization of Machine-Learning Models to Accurately Predict the Risk for Critical COVID-19," *Internal and Emergency Medicine* (15:8), pp. 1435–1443. (https://doi.org/10.1007/s11739-020-02475-0).

Aziz, S., Arabi, Y. M., Alhazzani, W., Evans, L., Citerio, G., Fischkoff, K., Salluh, J., Meyfroidt, G., Alshamsi, F., Oczkowski, S., Azoulay, E., Price, A., Burry, L., Dzierba, A., Benintende, A., Morgan, J., Grasselli, G., Rhodes, A., Møller, M. H., Chu, L., Schwedhelm, S., Lowe, J. J., Bin, D., and Christian, M. D. 2020. "Managing ICU Surge during the COVID-19 Crisis: Rapid Guidelines," *Intensive Care Medicine* (46:7), pp. 1303–1325. (https://doi.org/10.1007/s00134-020-06092-5).

Bai, X., Eyob, E., Ola, A., and Reese, S. 2021. "Factors Affecting Students' Satisfaction with Synchronous Online Instruction During the COVID-19 Pandemic," *Journal of International Technology and Information Management* (30:3), pp. 78–107.

Barstugan, M., Ozkaya, U., and Ozturk, S. 2020. "Coronavirus (Covid-19) Classification Using Ct Images by Machine Learning Methods," *ArXiv Preprint ArXiv:2003.09424*.

Bekker, R., and Koole, G. 2021. "Modeling COVID-19 Hospital Admissions and Occupancy in the Netherlands," *ArXiv Preprint ArXiv:2102.11021*.

Berta, P., Paruolo, P., Verzillo, S., and Lovaglio, P. G. 2020. "A Bivariate Prediction Approach for Adapting the Health Care System Response to the Spread of COVID-19," *PLOS ONE* (15:10), (T. Mathes, ed.), p. e0240150. (https://doi.org/10.1371/journal.pone.0240150).

Bonnasse-Gahot, L., Dénès, M., Dulac-Arnold, G., Girgin, S., Husson, F., Iovene, V., Josse, J., Kimmoun, A., Landes, F., and Nadal, J.-P. 2020. *ICU Bed Availability Monitoring and Analysis in the Grand Est Region of France during the COVID-19 Epidemic*.

Burdick, H., Lam, C., Mataraso, S., Siefkas, A., Braden, G., Dellinger, R. P., McCoy, A., Vincent, J.-L., Green-Saxena, A., Barnes, G., Hoffman, J., Calvert, J., Pellegrini, E., and Das, R. 2020. "Prediction of Respiratory Decompensation in Covid-19 Patients Using Machine Learning: The READY Trial," *Computers in*

Biology and Medicine (124), p. 103949. (https://doi.org/10.1016/j.compbiomed.2020.103949).

Buuren, S. van, and Groothuis-Oudshoorn, K. 2010. "Mice: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software*, University of California, Los Angeles, pp. 1–68.

Català, M., Alonso, S., Alvarez-Lacalle, E., López, D., Cardona, P.-J., and Prats, C. 2020. "Empirical Model for Short-Time Prediction of COVID-19 Spreading," *PLOS Computational Biology* (16:12), (B. M. Althouse, ed.), p. e1008431. (https://doi.org/10.1371/journal.pcbi.1008431).

CDC. 2020. "Coronavirus Disease 2019 (COVID-19)," *Centers for Disease Control and Prevention,* , February 11. (https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html, accessed December 8, 2021).

Chassagnon, G., Vakalopoulou, M., Battistella, E., Christodoulidis, S., Hoang-Thi, T.-N., Dangeard, S., Deutsch, E., Andre, F., Guillo, E., and Halm, N. 2020. "AI-Driven CT-Based Quantification, Staging and Short-Term Outcome Prediction of COVID-19 Pneumonia," *ArXiv Preprint ArXiv:2004.12852*.

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. 2018. "Recurrent Neural Networks for Multivariate Time Series with Missing Values," *Scientific Reports* (8:1), p. 6085. (https://doi.org/10.1038/s41598-018-24271-9).

Chen, T., and Guestrin, C. 2016. "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. (https://doi.org/10.1145/2939672.2939785).

Chin, V., Samia, N. I., Marchant, R., Rosen, O., Ioannidis, J. P., Tanner, M. A., and Cripps, S. 2020. "A Case Study in Model Failure? COVID-19 Daily Deaths and ICU Bed Utilisation Predictions in New York State," *European Journal of Epidemiology* (35:8), Springer, pp. 733–742.

Choudrie, J., Patil, S., Kotecha, K., Matta, N., and Pappas, I. 2021. "Applying and Understanding an Advanced, Novel Deep Learning Approach: A Covid 19, Text Based, Emotions Analysis Study," *Information Systems Frontiers* (23:6), pp. 1431–1465. (https://doi.org/10.1007/s10796-021-10152-6).

De Boor, C., and De Boor, C. 1978. *A Practical Guide to Splines*, (Vol. 27), springer-verlag New York.

Dhaliwal, S., Nahid, A.-A., and Abbas, R. 2018. “Effective Intrusion Detection System Using XGBoost,” *Information* (9:7), p. 149. (https://doi.org/10.3390/info9070149).

Di Castelnuovo, A., Bonaccio, M., Costanzo, S., Gialluisi, A., Antinori, A., Berselli, N., Blandi, L., Bruno, R., Cauda, R., Guaraldi, G., My, I., Menicanti, L., Parruti, G., Patti, G., Perlini, S., Santilli, F., Signorelli, C., Stefanini, G. G., Vergori, A., Abdeddaim, A., Ageno, W., Agodi, A., Agostoni, P., Aiello, L., Al Moghazi, S., Aucella, F., Barbieri, G., Bartoloni, A., Bologna, C., Bonfanti, P., Brancati, S., Cacciatore, F., Caiano, L., Cannata, F., Carrozzi, L., Cascio, A., Cingolani, A., Cipollone, F., Colomba, C., Crisetti, A., Crosta, F., Danzi, G. B., D’Ardes, D., de Gaetano Donati, K., Di Gennaro, F., Di Palma, G., Di Tano, G., Fantoni, M., Filippini, T., Fioretto, P., Fusco, F. M., Gentile, I., Grisafi, L., Guarnieri, G., Landi, F., Larizza, G., Leone, A., Maccagni, G., Maccarella, S., Mapelli, M., Maragna, R., Marcucci, R., Maresca, G., Marotta, C., Marra, L., Mastroianni, F., Mengozzi, A., Menichetti, F., Milic, J., Murri, R., Montineri, A., Mussinelli, R., Mussini, C., Musso, M., Odone, A., Olivieri, M., Pasi, E., Petri, F., Pinchera, B., Pivato, C. A., Pizzi, R., Poletti, V., Raffaelli, F., Ravaglia, C., Righetti, G., Rognoni, A., Rossato, M., Rossi, M., Sabena, A., Salinaro, F., Sangiovanni, V., Sanrocco, C., Scarafino, A., Scorzolini, L., Sgariglia, R., Simeone, P. G., Spinoni, E., Torti, C., Trecarichi, E. M., Vezzani, F., Veronesi, G., Vettor, R., Vianello, A., Vinceti, M., De Caterina, R., and Iacoviello, L. 2020. “Common Cardiovascular Risk Factors and In-Hospital Mortality in 3,894 Patients with COVID-19: Survival Analysis and Machine Learning-Based Findings from the Multicentre Italian CORIST Study,” *Nutrition, Metabolism and Cardiovascular Diseases* (30:11), pp. 1899–1913. (https://doi.org/10.1016/j.numecd.2020.07.031).

Elaziz, M. A., Hosny, K. M., Salah, A., Darwish, M. M., Lu, S., and Sahlol, A. T. 2020. “New Machine Learning Method for Image-Based Diagnosis of COVID-19,” *PLOS ONE* (15:6), (R. Damasevicius, ed.), p. e0235187. (https://doi.org/10.1371/journal.pone.0235187).

El-Ghitany, E. 2020. “A Short-Term Forecast Scenario for COVID-19 Epidemic and Allocated Hospital Readiness in Egypt,” preprint, Preprint, MEDICINE & PHARMACOLOGY, April 27. (https://doi.org/10.20944/preprints202004.0473.v1).

Farcomeni, A., Maruotti, A., Divino, F., Jona-Lasinio, G., and Lovison, G. 2021. “An Ensemble Approach to Short-term Forecast of COVID-19 Intensive Care Occupancy in Italian Regions,” *Biometrical Journal* (63:3), pp. 503–513. (https://doi.org/10.1002/bimj.202000189).

 ISSN: 1941-6679-On-line Copy

Francone, M., Iafrate, F., Masci, G. M., Coco, S., Cilia, F., Manganaro, L., Panebianco, V., Andreoli, C., Colaiacomo, M. C., Zingaropoli, M. A., Ciardi, M. R., Mastroianni, C. M., Pugliese, F., Alessandri, F., Turriziani, O., Ricci, P., and Catalano, C. 2020. "Chest CT Score in COVID-19 Patients: Correlation with Disease Severity and Short-Term Prognosis," *European Radiology* (30:12), pp. 6808–6817. (https://doi.org/10.1007/s00330-020-07033-y).

Friedman, J. H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine.," *The Annals of Statistics* (29:5). (https://doi.org/10.1214/aos/1013203451).

Funk, S., Abbott, S., Atkins, B., Baguelin, M., Baillie, J., Birrell, P., Blake, J., Bosse, N., Burton, J., Carruthers, J., Davies, N., De Angelis, D., Dyson, L., Edmunds, W., Eggo, R., Ferguson, N., Gaythorpe, K., Gorsich, E., Guyver-Fletcher, G., Hellewell, J., Hill, E., Holmes, A., House, T., Jewell, C., Jit, M., Jombart, T., Joshi, I., Keeling, M., Kendall, E., Knock, E., Kucharski, A., Lythgoe, K., Meakin, S., Munday, J., Openshaw, P., Overton, C., Pagani, F., Pearson, J., Perez-Guzman, P., Pellis, L., Scarabel, F., Semple, M., Sherratt, K., Tang, M., Tildesley, M., Van Leeuwen, E., Whittles, L., CMMID COVID-19 Working Group, Imperial College COVID-19 Response Team, and ISARIC4C Investigators. 2020. "Short-Term Forecasts to Inform the Response to the Covid-19 Epidemic in the UK," preprint, Preprint, Epidemiology, November 13. (https://doi.org/10.1101/2020.11.11.20220962).

Gao, Y., Cai, G.-Y., Fang, W., Li, H.-Y., Wang, S.-Y., Chen, L., Yu, Y., Liu, D., Xu, S., Cui, P.-F., Zeng, S.-Q., Feng, X.-X., Yu, R.-D., Wang, Y., Yuan, Y., Jiao, X.-F., Chi, J.-H., Liu, J.-H., Li, R.-Y., Zheng, X., Song, C.-Y., Jin, N., Gong, W.-J., Liu, X.-Y., Huang, L., Tian, X., Li, L., Xing, H., Ma, D., Li, C.-R., Ye, F., and Gao, Q.-L. 2020. "Machine Learning Based Early Warning System Enables Accurate Mortality Risk Prediction for COVID-19," *Nature Communications* (11:1), p. 5033. (https://doi.org/10.1038/s41467-020-18684-2).

García-Laencina, P. J., Sancho-Gómez, J.-L., and Figueiras-Vidal, A. R. 2010. "Pattern Classification with Missing Data: A Review," *Neural Computing and Applications* (19:2), Springer, pp. 263–282.

Goic, M., Bozanic-Leal, M. S., Badal, M., and Basso, L. J. 2021. "COVID-19: Short-Term Forecast of ICU Beds in Times of Crisis," *PLOS ONE* (16:1), (C.-H. Chen, ed.), p. e0245272. (https://doi.org/10.1371/journal.pone.0245272).

Gomes, D. P. S., Ulhaq, A., Paul, M., Horry, M. J., Chakraborty, S., Saha, M., Debnath, T., and Rahaman, D. M. M. 2020. "Potential Features of ICU Admission in X-Ray Images of COVID-19 Patients," *ArXiv:2009.12597 [Cs, Eess]*. (http://arxiv.org/abs/2009.12597).

Gumus, M., and Kiran, M. S. 2017. "Crude Oil Price Forecasting Using XGBoost," in *2017 International Conference on Computer Science and Engineering (UBMK)*, IEEE, pp. 1100–1103.

Haase, N., Plovsing, R., Christensen, S., Poulsen, L. M., Brøchner, A. C., Rasmussen, B. S., Helleberg, M., Jensen, J. U. S., Andersen, L. P. K., and Siegel, H. 2020. "Characteristics, Interventions, and Longer Term Outcomes of COVID-19 ICU Patients in Denmark—A Nationwide, Observational Study," *Acta Anaesthesiologica Scandinavica* (65:1), Wiley Online Library, pp. 68–75.

He, F., Quan, Y., Lei, M., Liu, R., Qin, S., Zeng, J., Zhao, Z., Yu, N., Yang, L., and Cao, J. 2020. "Clinical Features and Risk Factors for ICU Admission in COVID-19 Patients with Cardiovascular Diseases," *Aging and Disease* (11:4), p. 763. (https://doi.org/10.14336/AD.2020.0622).

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., and Gu, X. 2020. "Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China," *The Lancet* (395:10223), Elsevier, pp. 497–506.

Kaur, H., Ahsaan, S. U., Alankar, B., and Chang, V. 2021. "A Proposed Sentiment Analysis Deep Learning Algorithm for Analyzing COVID-19 Tweets," *Information Systems Frontiers* (23:6), pp. 1417–1429. (https://doi.org/10.1007/s10796-021-10135-7).

Keeling, M. J., Dyson, L., Guyver-Fletcher, G., Holmes, A., Semple, M. G., ISARIC4C Investigators, Tildesley, M. J., and Hill, E. M. 2020. "Fitting to the UK COVID-19 Outbreak, Short-Term Forecasts and Estimating the Reproductive Number," preprint, Preprint, Infectious Diseases (except HIV/AIDS), August 4. (https://doi.org/10.1101/2020.08.04.20163782).

Kreindler, D. M., and Lumsden, C. J. 2012. "The Effects of the Irregular Sample and Missing Data in Time Series Analysis," *Nonlinear Dynamical Systems Analysis for the Behavioral Sciences Using Real Data* (135:2).

Lipton, Z. C., Kale, D. C., and Wetzel, R. 2016. "Modeling Missing Data in Clinical Time Series with RNNs," *ArXiv:1606.04130 [Cs, Stat]*. (http://arxiv.org/abs/1606.04130).

 ISSN: 1941-6679-On-line Copy

Manevski, D., Ružić Gorenjec, N., Kejžar, N., and Blagus, R. 2020. "Modeling COVID-19 Pandemic Using Bayesian Analysis with Application to Slovene Data," *Mathematical Biosciences* (329), p. 108466. (https://doi.org/10.1016/j.mbs.2020.108466).

Massaoudi, M., Refaat, S. S., Abu-Rub, H., Chihi, I., and Wesleti, F. S. 2020. "A Hybrid Bayesian Ridge Regression-CWT-Catboost Model For PV Power Forecasting," in *2020 IEEE Kansas Power and Energy Conference (KPEC)*, Manhattan, KS, USA: IEEE, July, pp. 1–5. (https://doi.org/10.1109/KPEC47870.2020.9167596).

Massonnaud, C., Roux, J., and Crépey, P. 2020. "COVID-19: Forecasting Short Term Hospital Needs in France," preprint, Preprint, Epidemiology, March 20. (https://doi.org/10.1101/2020.03.16.20036939).

Mostafa, S. M., Eladimy, A. S., Hamad, S., and Amano, H. 2020. "CBRG: A Novel Algorithm for Handling Missing Data Using Bayesian Ridge Regression and Feature Selection Based on Gain Ratio," *IEEE Access* (8), pp. 216969–216985. (https://doi.org/10.1109/ACCESS.2020.3042119).

Nobre, J., and Neves, R. F. 2019. "Combining Principal Component Analysis, Discrete Wavelet Transform and XGBoost to Trade in the Financial Markets," *Expert Systems with Applications* (125), pp. 181–194. (https://doi.org/10.1016/j.eswa.2019.01.083).

Ogunleye, A., and Wang, Q.-G. 2020. "XGBoost Model for Chronic Kidney Disease Diagnosis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (17:6), pp. 2131–2140. (https://doi.org/10.1109/TCBB.2019.2911071).

Ozturk, T., Talo, M., Yildirim, E. A., Baloglu, U. B., Yildirim, O., and Rajendra Acharya, U. 2020. "Automated Detection of COVID-19 Cases Using Deep Neural Networks with X-Ray Images," *Computers in Biology and Medicine* (121), p. 103792. (https://doi.org/10.1016/j.compbiomed.2020.103792).

Petermann, M., Frhlich, J., and Wyler, D. 2021. "A Case Study of Short-Term Forecasting during the COVID-19 Pandemic: The Platform Icumonitoring.Ch," *Swiss Medical Weekly*. (https://doi.org/10.4414/smw.2021.20524).

Rácz, A., Bajusz, D., and Héberger, K. 2021. *Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification*, p. 16.

 ISSN: 1941-6679-On-line Copy

Rahaman Khan, M. H., and Hossain, A. 2020. "COVID-19 Outbreak Situations in Bangladesh: An Empirical Analysis," preprint, Preprint, Infectious Diseases (except HIV/AIDS), April 22. (https://doi.org/10.1101/2020.04.16.20068312).

Ricoca Peixoto, V., Vieira, A., Aguiar, P., Carvalho, C., Thomas, D., and Abrantes, A. 2020. "Rapid Assessment of the Impact of Lockdown on the COVID-19 Epidemic in Portugal," preprint, Preprint, Epidemiology, May 27. (https://doi.org/10.1101/2020.05.26.20098244).

Roncon, L., Zuin, M., Rigatelli, G., and Zuliani, G. 2020. "Diabetic Patients with COVID-19 Infection Are at Higher Risk of ICU Admission and Poor Short-Term Outcome," *Journal of Clinical Virology* (127), p. 104354. (https://doi.org/10.1016/j.jcv.2020.104354).

Rubin, D. B. 1976. "Inference and Missing Data," *Biometrika* (63:3), Oxford University Press, pp. 581–592.

Rubin, D. B. 1996. "Multiple Imputation after 18+ Years," *Journal of the American Statistical Association* (91:434), Taylor & Francis Group, pp. 473–489.

Schumaker, R. P., Veronin, M. A., Rohm, T., Boyett, M., and Dixit, R. R. 2021. "A Data Driven Approach to Profile Potential SARS-CoV-2 Drug Interactions Using TylerADE," *Journal of International Technology and Information Management* (30:3), pp. 108–142.

Schumaker, R., Veronin, M., Rohm, T., Dixit, R., Aljawarneh, S., and Lara, J. 2021. "An Analysis of Covid-19 Vaccine Allergic Reactions," *Journal of International Technology and Information Management* (30:4), pp. 24–40.

Singh, P. D., Kaur, R., Singh, K. D., and Dhiman, G. 2021. "A Novel Ensemble-Based Classifier for Detecting the COVID-19 Disease for Infected Patients," *Information Systems Frontiers* (23:6), pp. 1385–1401. (https://doi.org/10.1007/s10796-021-10132-w).

Solimando, A. G., Susca, N., Borrelli, P., Prete, M., Lauletta, G., Pappagallo, F., Buono, R., Inglese, G., Forina, B. M., Bochicchio, D., Capobianco, M., Carrieri, V., Cicco, S., Leone, P., Silvestris, N., Saracino, A., Ria, R., Procacci, V., Migliore, G., Vacca, A., and Racanelli, V. 2021. "Short-Term Variations in Neutrophil-to-Lymphocyte and Urea-to-Creatinine Ratios Anticipate Intensive Care Unit Admission of COVID-19 Patients in the Emergency Department," *Frontiers in Medicine* (7), p. 625176. (https://doi.org/10.3389/fmed.2020.625176).

Song, Q., Shepperd, M., Chen, X., and Liu, J. 2008. "Can K-NN Imputation Improve the Performance of C4. 5 with Small Software Project Data Sets? A

Comparative Evaluation," *Journal of Systems and Software* (81:12), Elsevier, pp. 2361–2370.

Sun, Liping, Song, F., Shi, N., Liu, F., Li, S., Li, P., Zhang, W., Jiang, X., Zhang, Y., Sun, Lining, Chen, X., and Shi, Y. 2020. "Combination of Four Clinical Indicators Predicts the Severe/Critical Symptom of Patients Infected COVID-19," *Journal of Clinical Virology* (128), p. 104431. (https://doi.org/10.1016/j.jcv.2020.104431).

Torlay, L., Perrone-Bertolotti, M., Thomas, E., and Baciu, M. 2017. "Machine Learning–XGBoost Analysis of Language Networks to Classify Patients with Epilepsy," *Brain Informatics* (4:3), pp. 159–169. (https://doi.org/10.1007/s40708-017-0065-7).

Vultaggio, A., Vivarelli, E., Virgili, G., Lucenteforte, E., Bartoloni, A., Nozzoli, C., Morettini, A., Berni, A., Malandrino, D., Rossi, O., Nencini, F., Pieralli, F., Peris, A., Lagi, F., Scocchera, G., Spinicci, M., Trotta, M., Mazzetti, M., Parronchi, P., Cosmi, L., Liotta, F., Fontanari, P., Mazzoni, A., Salvati, L., Maggi, E., Annunziato, F., Almerigogna, F., and Matucci, A. 2020. "Prompt Predicting of Early Clinical Deterioration of Moderate-to-Severe COVID-19 Patients: Usefulness of a Combined Score Using IL-6 in a Preliminary Study," *The Journal of Allergy and Clinical Immunology: In Practice* (8:8), pp. 2575-2581.e2. (https://doi.org/10.1016/j.jaip.2020.06.013).

Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., Wang, B., Xiang, H., Cheng, Z., and Xiong, Y. 2020. "Clinical Characteristics of 138 Hospitalized Patients with 2019 Novel Coronavirus–Infected Pneumonia in Wuhan, China," *Jama* (323:11), American Medical Association, pp. 1061–1069.

Weissman, G. E., Crane-Droesch, A., Chivers, C., Luong, T., Hanish, A., Levy, M. Z., Lubken, J., Becker, M., Draugelis, M. E., Anesi, G. L., Brennan, P. J., Christie, J. D., Hanson, C. W., Mikkelsen, M. E., and Halpern, S. D. 2020. "Locally Informed Simulation to Predict Hospital Capacity Needs During the COVID-19 Pandemic," *Annals of Internal Medicine* (173:1), pp. 21–28. (https://doi.org/10.7326/M20-1260).

Wu, J., Zhang, P., Zhang, L., Meng, W., Li, J., Tong, C., Li, Y., Cai, J., Yang, Z., Zhu, J., Zhao, M., Huang, H., Xie, X., and Li, S. 2020. "Rapid and Accurate Identification of COVID-19 Infection through Machine Learning Based on Clinical Available Blood Test Results," preprint, Preprint, Infectious Diseases (except HIV/AIDS), April 6. (https://doi.org/10.1101/2020.04.02.20051136).

 ISSN: 1941-6679-On-line Copy

Zhao, C., Tepekule, B., Criscuolo, N. G., Wendel Garcia, P. D., Hilty, M. P., RISC-19-ICU consortium Investigators in Switzerland, Fumeaux, T., and Van Boeckel, T. 2020. "Icumonitoring.Ch: A Platform for Short-Term Forecasting of Intensive Care Unit Occupancy during the COVID-19 Epidemic in Switzerland," *Swiss Medical Weekly*. (https://doi.org/10.4414/smw.2020.20277).

Zhao, H., Merchant, N. N., McNulty, A., Radcliff, T. A., Cote, M. J., Fischer, R. S. B., Sang, H., and Ory, M. G. 2021. "COVID-19: Short Term Prediction Model Using Daily Incidence Data," *PLOS ONE* (16:4), (J. Schieffelin, ed.), p. e0250110. (https://doi.org/10.1371/journal.pone.0250110).

Zhao, Z., Chen, A., Hou, W., Graham, J. M., Li, H., Richman, P. S., Thode, H. C., Singer, A. J., and Duong, T. Q. 2020. "Prediction Model and Risk Scores of ICU Admission and Mortality in COVID-19," *PLOS ONE* (15:7), (M. Adrish, ed.), p. e0236618. (https://doi.org/10.1371/journal.pone.0236618).

 ISSN: 1941-6679-On-line Copy