Hindawi
Scientific Programming
Volume 2021, Article ID 5587188, 10 pages
https://doi.org/10.1155/2021/5587188

Research Article

Machine Learning-Based Model to Predict the Disease Severity and Outcome in COVID-19 Patients

Sumayh S. Aljameel,[1] **Irfan Ullah Khan**,[1] **Nida Aslam**,[1] **Malak Aljabri**,[1] **and Eman S. Alsulmi**[2]

[1]*College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia*
[2]*Department of Obstetrics and Gynecology, College of Medicine, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia*

Correspondence should be addressed to Sumayh S. Aljameel; saljameel@iau.edu.sa

Received 1 February 2021; Revised 6 March 2021; Accepted 10 April 2021; Published 20 April 2021

Academic Editor: Shah Nazir

The novel coronavirus (COVID-19) outbreak produced devastating effects on the global economy and the health of entire communities. Although the COVID-19 survival rate is high, the number of severe cases that result in death is increasing daily. A timely prediction of at-risk patients of COVID-19 with precautionary measures is expected to increase the survival rate of patients and reduce the fatality rate. This research provides a prediction method for the early identification of COVID-19 patient's outcome based on patients' characteristics monitored at home, while in quarantine. The study was performed using 287 COVID-19 samples of patients from the King Fahad University Hospital, Saudi Arabia. The data were analyzed using three classification algorithms, namely, logistic regression (LR), random forest (RF), and extreme gradient boosting (XGB). Initially, the data were preprocessed using several preprocessing techniques. Furthermore, 10-k cross-validation was applied for data partitioning and SMOTE for alleviating the data imbalance. Experiments were performed using twenty clinical features, identified as significant for predicting the survival versus the deceased COVID-19 patients. The results showed that RF outperformed the other classifiers with an accuracy of 0.95 and area under curve (AUC) of 0.99. The proposed model can assist the decision-making and health care professional by early identification of at-risk COVID-19 patients effectively.

1. Introduction

Coronavirus (COVID-19) started in China in December 2019. As of January 2021, over 95 million cases have been reported around the world, with a mortality rate of 2% of the total closed cases [1]. This rapid pandemic expansion represents a global concern and a serious threat to the public health and economy worldwide. To prevent the infection from spreading, most countries restricted social interaction through precautionary measures such as isolation and quarantine. However, many infected patients did not benefit from the proper treatment due to late diagnosis and the novel and unknown nature of the virus. Recently, many researchers focused on developing new methodologies to screen infected patients in different stages to find notable associations between the patient's clinical features and the chances to succumb to the disease [2, 3]. Current investigation studies determined that artificial intelligence (AI) and machine learning (ML) techniques can play a key role in reducing the effect of the virus spread [4–6]. ML application technologies on patients' data fall under a range of different research directions [7]. One of the most important research directions is predicting the infection rate and mortality rate and building a model to classify patients based on their clinical findings [8, 9]. These research investigations are extremely important and would greatly assist people in the health sectors to be well prepared and take all necessary precautions to minimize the pandemic spread.

The aim of this research is to develop a prediction model to calculate the severity of the disease in COVID-19 patients, using risk factors that can be monitored remotely, with the patient being at home. Moreover, the study explores the impact of vital signs, chronic diseases, preliminary clinical

investigations, and demographic features to predict the survival versus the mortality of COVID-19 patients. The study used COVID-19 patients' data from the King Fahad University Hospital containing the clinical findings and demographic information to validate the model performance and effectiveness. All the risk factors or vital signs that can be measured through widely used sensors were included in the study such as oxygen level in the blood, temperature, pulse rate, and blood pressure. The model will serve as an early warning system to timely identify at-risk patients.

1.1. Related Work. Early detection and diagnosis using AI techniques help to prevent the spread and to combat the COVID-19 pandemic using different data such as CT scans, X-ray, clinical data, and blood sample data.

Yan et al. [10] predicted the criticality and survival chances of patients with severe COVID-19 infection based on different risk factors and demographic information. The dataset used consists of 375 records from patients admitted to Tongji Hospital from January 10th to February 18th, 2020, including 201 survivors and 174 deceased within the same period. They used an XGBoost (XGB) model and identified only three main clinical features as significant, i.e., lactic dehydrogenase (LDH), lymphocyte, and high-sensitivity C-reactive protein (Hs-CRP), selected from more than 300 features. The proposed model was validated using data from 29 patients. The key findings of the research were the model's ability to predict the risk of death with 0.95 precision and 0.90 prediction accuracy. Such models will equip physicians with a tool for identifying critical conditions, thereby helping to reduce the mortality rate. Even though these findings are of great importance, the research has some limitations, which affect the accuracy of the reported results. These limitations were due to the small size of the dataset, namely, 29 records of patients only.

Similarly, Wong and So [11] also used XGB with another dataset to predict the severe and the death cases and identify the risk factors associated with COVID-19. The dataset was retrieved from United Kingdom Biobank (UKBB) and includes 93 different variables collected between 16 March 2020 and 19 July 2020. Two different studies have been conducted based on the sample's groups. For the first study, the data were clinical prediagnostic data of 1747 COVID-19 infected patient records containing both severe and death cases. For the severity class, the accuracy achieved was 0.668, and for the fatality class, the accuracy was 0.712. For the second study, the data were taken from the negative cases, the general population with no COVID-19 infection, consisting of 489987 records. The same model was applied, and the accuracy achieved was similar to the first study, with an accuracy of 0.669 for the severity class and 0.749 for the fatality class, respectively. It is worth mentioning that the researchers identified the five most significant risk factors for severe cases and death cases, with age being the top factor for both cases. Other factors include obesity, impaired renal function, multiple comorbidities, and cardiometabolic abnormalities.

Sun et al. [12] developed a prediction model using the support vector machine (SVM) to predict the severe cases of COVID-19 patients. In the study, they used the clinical and laboratory features that are significantly associated with these cases. Using 336 cases of COVID-19 patients, 26 severe/critical cases and 310 noncritical, they found that the main features to discriminate the mild and severe cases are age, growth hormone secretagogues (GHSs), immune feature cluster of differentiation 3 (CD3) percentage, and total protein. They found that the proposed model was effective and robust in predicting patients in severe conditions with up to 0.775 accuracy.

Another research conducted by Yao et al. [13] also applied the SVM model to classify the COVID-19 patients according to the severity of the symptoms. They applied SVM for the binary class label on a total of 137 records including urine and blood test results and combining both severely ill patients and patients with mild symptoms. The results showed that around 32 factors have high correlations with severe COVID-19, with an accuracy of 0.815. It is worth mentioning that, amongst all factors, age and gender had mostly affected the classification of cases between severe and mild. Patients aged around 65 had more severe cases than others. Moreover, male patients were at a higher risk of developing severe COVID-19 symptoms. In terms of the urine and blood test samples, blood test result features show more significant differences between severe and mild cases than urine test result features.

Hu et al. [14] used the logistic regression (LR) model to identify the COVID-19 patients' severity. They used a dataset containing demographic and clinical data for 115 COVID-19 patients under the nonsevere condition and 68 COVID-19 patients under the severe condition. Four features have been selected as the most significant features to discriminate the mild and severe cases: age, high-sensitivity C-reactive protein level, lymphocyte count, and d-dimer level. This model was evaluated, and the results showed that the prediction was effective with area under the receiver operating characteristic (AUROC) of 0.881, sensitivity of 0.839, and specificity of 0.794, respectively. Bertsimas et al. [15] used 3927 COVID-19 patients' sample for predicting the mortality risk using XGB. The study used demographic and the clinical features of the patients from 33 hospital data. The model achieved the accuracy of 0.85 and AUC of 0.90. Moreover, Sánchez-Montañés et al. [16] developed LR-based mortality prediction using 1969 COVID-19-positive patients. The study found age and O_2 as the significant features and achieved an AUC of 0.89, sensitivity of 0.82, and specificity of 0.81, respectively.

In [5], supervised machine learning techniques have been investigated to predict the COVID-19 outbreak. In [5], SVM has been used for prediction over the dataset obtained from the WHO with 303 patients. The proposed scheme exhibits an accuracy of 0.967 during the testing phase. Similarly, An et al. [17] developed the model to predict the mortality of COVID-19 patients using several machine learning algorithms such as LASSO, SVM (linear and RBF), RF, and KNN. The models were trained to identify three cases, i.e., mortality and survived and mortality and survived

within 14 and 30 days after the initial diagnosis. Linear SVM achieved the highest performance with an AUC of 0.962, sensitivity of 0.92, and specificity of 0.91, respectively. The study found age, diabetes mellitus, and cancer as a significant factor in the mortality prediction for COVID-19 patients.

In conclusion, the importance of machine learning specifically, on predictive analysis, has been proven from several studies. Some of the studies have been conducted to perform the prediction and forecasting, yet there is still a need for further exploration and to extend the findings associated with COVID-19 using a real dataset of clinical records. The summary of the related studies is shown in Table 1. The proposed model in this study attempts to predict and forecast the patients that are at risk along with identifying the main risk factors associated with COVID-19. Targeted patients are isolated at home. The dataset (clinical findings) has been retrieved from King Fahad University Hospital in the Kingdom of Saudi Arabia. The main aim of the study is to develop a preemptive warning model that can identify at-risk COVID-19 patients that are monitored in quarantine at home.

This paper is organized as follows: Section 2 introduces the materials and methods, and Section 3 shows the experimental setup and results. Finally, the conclusion and future work are identified in Section 4.

2. Methodology

The following section covers the dataset description and the methodology used. Due to the class imbalance in the dataset, the synthetic minority oversampling technique (SMOTE) was used.

2.1. Dataset Description. The study was conducted in the Department of Computer Science of Imam Abdulrahman bin Faisal University (IAU) and approved by the Deanship of Scientific Research of IAU under the research grant IRB-2020-09-160. The data were collected from King Fahad University Hospital, Dammam, Kingdom of Saudi Arabia (KSA). The dataset contains the demographic and clinical data of COVID-19-positive patients in the period from 30 April 2020 to 24 July 2020. The dataset contains all the positive patients that were admitted in King Fahad University Hospital during the specified data collection period. There are 287 COVID-19 patient records in the dataset with a binary class label, namely, "survived" and "deceased," respectively. The number of survived patients is 243, and 44 patients deceased. The distribution of instances per class label is shown in Figure 1, while the description of the dataset is mentioned in Table 2. The field BodyTemp 1 in the table indicates the first body temperature taken at the time of the patient's admission to the hospital. However, BodyTemp 2 indicates the last body temperature reading taken before the patient's discharge. Similarly, SOB indicates shortness of breath, chr_dm indicates chronic disease diabetes mellitus, chr_htn indicates hypertension, chr_cardiac represents cardiovascular diseases, chr_dlp represents dyslipidemia, and chr_ckd indicates chronic kidney disease.

The baseline characteristics of the numeric attributes of the dataset are represented in terms of mean ± standard deviation (SD). By contrast, the categorical attributes are measured by a count. The characteristics of the features in the dataset are presented in Table 3.

2.2. Preprocessing. Preprocessing is one of the key steps in data analysis and prediction. Several preprocessing techniques were applied on the dataset. The dataset contains data of all the patients admitted in the hospital. Some symptoms or vital signs occurred with very low frequency and were therefore removed from the dataset. All symptoms with occurrences at 50% or above were selected to be added to the feature set, while the symptoms with occurrences in the range from 2% to 49% were cumulated as one feature the was assigned a unique code. The first three vital signs: fever, cough, and shortness of breath (SOB) were defined as symptom features, while the remaining features were incorporated as a new attribute "sym_others." 5% of the patients in the study were asymptomatic at the time of initial diagnosis and considered as a part of the sym_others attribute. Similarly, the chronic top three (3) diseases (i.e., diabetes, high blood pressure, and cardiac) with the highest frequency were included as features. However, all other chronic disease types with more than 1 occurrence were incorporated as one feature "chr_others." After the initial preprocessing data, an encoding scheme was applied on the categorical features. As the dataset contains a small number of missing values, imputation was performed using the *K*-means technique.

2.3. Prediction Model. In the study, three classification algorithms were used: logistic regression (LR), random forest, and extreme gradient boosting (XGB). A brief description of the classification algorithms is given below.

2.3.1. Logistic Regression. Logistic regression is one of the widely used statistical classification algorithms for binary and multiclass problems. For predicting the probability of the class label, logistic function is used [18]. The functional form of the hypothesis is

$$\mathbf{Y} = \mathbf{C}^{\mathbf{T}}(\mathbf{X}), \tag{1}$$

where C is the list of regression coefficients and X is the list of the features.

$$\mathbf{C} = \begin{bmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \cdots \\ \boldsymbol{\beta}_{\mathbf{n}} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \mathbf{X}_3 \\ \cdots \\ \mathbf{X}_{\mathbf{n}} \end{bmatrix}, \tag{2}$$

where $\beta_{\mathbf{i}}$ represents the regression estimators also known as predicted weights for the selected features in the data and β_0 represents the intercept of the equation.

Table 1: Related studies on mortality prediction for COVID-19 patients.

Reference	Technique	Dataset	Target class	Result
[10]	XGB	404 patients	Death, survived	0.95 precision 0.90 accuracy
[11]	XGB	1747 COVID-19 patients	Fatal, severe	Accuracy 0.668 (fatality) 0.712 (severe)
[12]	SVM	336 COVID-19 patients	Severe, critical	0.775 accuracy
[13]	SVM	137 COVID-19 patients	Severe, nonsevere	0.815 accuracy
[14]	LR	115 COVID-19 patients	Severe, nonsevere	0.881 AUROC 0.839 sensitivity 0.794 specificity
[5]	SVM	303 patients	Negative, positive cases	0.967 accuracy
[15]	XGB	3927 COVID-19 patients	—	0.85 accuracy 0.90 AUC
[16]	LR	1696 COVID-19 patients	Home, deceased	0.89 AUC 0.82 sensitivity 0.81 specificity
[17]	SVM (linear)	8000 COVID-19 patients	Mortality, recovered	0.962 AUC 0.92 sensitivity 0.91 specificity

Figure 1: Number of records per class label.

$$H(x) = Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n. \quad (3)$$

Since the dataset used in the study consists of 25 features in total, the logistic regression algorithm for our study is

$$\mathbf{h}(\mathbf{x}) = (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n). \quad (4)$$

The model will predict the record as survived or death if the value of

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \geq 0. \quad (5)$$

For optimal selection of regression estimator, maximum-likelihood ratio concept is used.

Sigmoid function (logistic function) is used to map the attributes with the class label. The functional form of the sigmoid equation is given in the following equations:

$$\mathbf{S}(\mathbf{g}) = \frac{1}{(1 + \mathbf{e}^{-y})}, \quad (6)$$

$$\mathbf{S}(\mathbf{g}) = \frac{1}{\left(1 + \mathbf{e}^{-C^T(X)}\right)}, \quad (7)$$

where e is a numeric constant Euler's number. In LR, a regularization parameter is used to reduce the chance of model overfitting. The logistic regression was optimized using grid search to get hyperoptimized parameters. The parameter set for logistic regression used in our study is shown in Table 4.

2.3.2. Random Forest. Random forest is an ensemble-based classification and regression model initially proposed by Zhang [19]. Random forest can be used for feature selection as well. It uses the bootstrapping data sampling method for partitioning of the data into training and testing sets. The model iteratively generates the trees for every bootstrap. The final prediction is made using the mean vote for each class. It is the combination of all generated decision trees. A decision tree is the hierarchical classification algorithm. The selection of the decision node is made using entropy, information gain, gain ratio, and Gini-index, respectively. In our study, we used information gain and entropy, as shown in the following equations:

$$\mathbf{E}(\mathbf{Y}) = \sum_{i=1}^{n} -\mathbf{p_i}\mathbf{log_2}\mathbf{p_i}, \quad (8)$$

$$\mathbf{E}(\mathbf{X}, \mathbf{Y}) = \sum_{n \in X} \mathbf{P}(\mathbf{n})\mathbf{E}(\mathbf{n}), \quad (9)$$

where $E(Y)$ represents the entropy of the target, while $\mathbf{Entropy}(\mathbf{X}, \mathbf{Y})$ is the entropy of the attributes with the target, in which $X = \{x_1, x_2, \ldots, x_n\}$ is the set of attributes in the dataset. The attribute with the highest information gain will be the root attribute, as follows:

$$\text{Information\_Gain} = E(Y) - E(X, Y). \quad (10)$$

It combines the predictions made by multiple trees using randomly selected vectors represented by θ_T. The selected

Table 2: Description of the dataset.

No.	Feature name	UOM	Data type	Missing values
1	Age	Years	Numeric	0
2	Gender	Male/female	Nominal	0
3	BodyTemp (1&2)	Celsius (°C)	Numeric	1%–11%
4	Pulse rate (1&2)	Beats per minute (BPM)	Numeric	7%–5%
5	Resp (1&2)	Breaths per minute (BPM)	Numeric	3%–3%
6	BP_Sys (1&2)	mm Hg	Numeric	10%–7%
7	BP_Dsys (1&2)	mm Hg	Numeric	5%–5%
8	OX (1&2)	mm Hg	Numeric	4%–5%
9	Fever	Yes/no	Nominal	0
10	SOB	Yes/no	Nominal	0
11	Cough	Yes/no	Nominal	0
12	Symptoms_Others	—	Nominal	0
13	chr_dm	Yes/no	Nominal	0
14	chr_htn	Yes/no	Nominal	0
15	chr_cardiac	Yes/no	Nominal	0
16	chr_dlp	Yes/no	Nominal	0
17	Chr_ckd	Yes/no	Nominal	0
18	Chr disease_others	—	Nominal	0

vectors are independent with the previously selected vectors. This results in the collection of trees represented by $h(x)$. The generalization error of decision tree is represented as follows:

$$\mathbf{GE} = \mathbf{P}_{\mathbf{X,Y}} \quad (\mathbf{margin\_fuc}(\mathbf{X}, \mathbf{Y}) < 0), \tag{11}$$

where $\mathbf{P}_{\mathbf{X,Y}}$ is the probability of set of the attributes to map to class label Y.

The parameters used in our study for random forest classifier are shown in Table 5.

2.3.3. Extreme Gradient Boosting. Extreme gradient boosting (XGB) algorithm is an ensemble-based classification and regression technique. It is the regularized form of the gradient boosting algorithm. Gradient boosting algorithm due to the data imbalance sometimes suffers from model overfitting. However, in the XGB algorithm, the regularization parameter reduces the risk the model overfitting. Like random forest, XGB is also a tree-based ensemble classifier. The boosting data resampling method attempts to enhance the model accuracy by minimizing the misclassification error [19]. It is an iterative approach. The records that were not successfully predicted in the previous iteration were used in the next iteration for training the model. The model will repeat the process until the model achieved an optimal result.

The regularization parameter reduces the variance in the model by increasing the weights of the misclassified instances. The increase in weight decreases the model underfitting. However, for reducing the bias of the model, penalty regularization was used to control the model overfitting without leading to a high misclassification rate. The XGB algorithm is the combination of several parameters. The optimal combination of parameters enhances the performance of the model. For parameter optimization, the gird search technique was used. The parameter used in the XGB algorithm is represented in Table 6.

2.4. Performance Evaluation. The performance of the model was evaluated using the standard evaluation measures such as accuracy, precision, sensitivity, specificity, and F-score, respectively. Area under curve and receiver operating characteristic (ROC) were also used for comparing the classifiers. It is one of the widely used tests for exploring the trade-off between true-positive (sensitivity) and false-positive rate (specificity) for the diagnostic test.

$$\mathbf{accuracy} = \frac{\mathbf{TP} + \mathbf{TN}}{\mathbf{TP} + \mathbf{TN} + \mathbf{FP} + \mathbf{FN}}, \tag{12}$$

where the accuracy of the model represents the proportion of the test records that is correctly classified.

$$\mathbf{Sensitivity} = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FN}}. \tag{13}$$

Sensitivity is the proportion of the positive class labels that is correctly predicted. It is also known as the true-positive rate (TPR) or positive-predicted value (PPV).

$$\mathbf{Specificity} = \frac{\mathbf{TN}}{\mathbf{TN} + \mathbf{FP}}. \tag{14}$$

Sensitivity also known as the true-negative rate (TNR) or negative-predicted value (NPV) is the proportion of the negative class labels that are correctly predicted as negative.

$$\mathbf{F-score} = \frac{2 \times \mathbf{precision} \times \mathbf{recall}}{\mathbf{precision} + \mathbf{recall}}, \tag{15}$$

where F-score is the harmonic mean of precision and recall.

3. Experimental Setup and Results

Data imbalance is one of the challenges in data analysis and usually leads to model overfitting. The dataset in this study also suffers from data imbalance as presented in Figure 1. The number of records for the survived category is 243 and for death category is 44. K-nearest neighbor-(KNN-) based synthetic minority oversampling

Table 3: Characteristics of the samples in the dataset.

Feature type	Feature name	Survived ($n = 243$)	Death ($n = 44$)
Demographic	Age (years)	47.28 ± 15.84	59.3 ± 14.4
	Gender		
	Male	159 (65)	38 (86)
	Female	86 (35)	8 (18)
Preliminary investigation	BodyTemp 1	37.1 ± 4.25	33.9 ± 10.7
	Pulse rate-1	94.6 ± 21.6	93.4 ± 31.1
	Resp-1	94.6 ± 21.6	93.4 ± 31.08
	BP_Sys-1	126.45 ± 21.52	128.7 ± 41.27
	BP_Dsys1	78.4 ± 14.9	75.5 ± 26.6
	OX1	93.2 ± 11.9	77.0 ± 24.8
	Temp2	14.8 ± 230.9	0 ± 17.6
	Pulse2	77.1 ± 28.9	68.5 ± 47.7
	Resp2	17.8 ± 6.6	17.8 ± 6.6
	BP_Sys2	109.7 ± 38.9	73.4 ± 55.8
	BP_Dsys2	68.0 ± 24.3	45.2 ± 35.1
	OX2	68.0 ± 24.4	45.2 ± 35.1
Symptoms		Fever	
	Yes	144 (59)	22 (50)
	No	99 (40)	22 (50)
		SOB	
	Yes	111 (45)	31 (70)
	No	134 (55)	13 (30)
		Cough	
	Yes	131 (53)	20 (45)
	No	114 (46)	24 (55)
		Other symptoms	
	Fatigue_weakness	16 (7)	4 (9)
	Sore_throat	10 (4)	3 (7)
	Pain	17 (7)	2 (5)
	Diar	12 (5)	1 (2)
	Anorexia	5 (2)	3 (7)
	Dizz	9 (4)	1 (2)
	Headache	11 (5)	3 (7)
	Nausea	13 (5)	2 (5)
	Vomit	8 (3)	1 (2)
	Dyspnea	13 (5)	2 (5)
	Runny_nose	6 (2)	1 (2)
	Chill	5 (2)	3 (7)
	No	118 (49)	32 (32)

Table 3: Continued.

Feature type	Feature name	Survived ($n = 243$)	Death ($n = 44$)
Chronic disease	chr_dm		
	Yes	73 (30)	23 (52)
	No	170 (69.9)	21 (48)
	chr_htn		
	Yes	67 (27)	18 (41)
	No	176 (72)	26 (59)
	chr_cardic		
	Yes	24 (9.8)	9 (20)
	No	219 (90)	35 (80)
	chr_dlp		
	Yes	25 (10.2)	3 (7)
	No	218 (89.7)	41 (93)
	Chr_CKD		
	Yes	16 (6.5)	5 (11)
	No	227 (93.4)	39 (39)
	Other_ChrDis		
	Epilepsy	4 (2)	1 (2)
	Stroke	4 (2)	3 (7)
	Respiratory	5 (2)	1 (2)
	Bph	3 (1)	2 (5)
	Sle	4 (2)	1 (2)
	Obesity	2 (1)	2 (5)
	Hypothyroidism	5 (2)	1 (2)
	Sickle	4 (2)	1 (2)
	Anemia	4 (2)	1 (2)
	Asthma	2 (1)	3 (7)
	Bone	4 (2)	2 (5)
	Ba	4 (2)	1 (2)
	Dyslipidemia	3 (1)	2 (5)
	Sinusitis	3 (1)	1 (2)
	Dpl	3 (1)	1 (2)
	No	189 (78)	25 (57)

Table 4: Logistic regression parameters using grid search optimization.

Parameter name	Value
Penalty	L2
Random_state	777
Max_iter	10000
Tol	10

Table 5: Random forest parameters using grid search optimization.

Parameter name	Value
Random_state	1
N_estimators	100
Max_depth	15
Min_samples_split	5
Min_samples_leaf	1

Table 6: XGB parameters using grid search optimization.

Parameter name	Value
learning_rate	0.05
max_depth	3
max_features	0.5
random_state	42

technique (SMOTE) was used to alleviate the data imbalance. SMOTE is an algorithm developed by Chawla et al. [20] to overcome the issue of imbalanced datasets in machine learning. In the SMOTE algorithm, the k-nearest neighbor (KNN) is used to calculate the Euclidean distance between the minority class instances to generate new minority class samples in the neighborhood. For A is the minority class with x instances, $A = \{x_1, x_2, \ldots x_n\}$ and k-nearest neighbors of $x_1 = \{x_6, x_7, \ldots x_k\}$ and then A_1 of $x_1 = \{x_7, x_4, \ldots x_n\}$, where $xk€A1$ ($k = 1, 2, 3, \ldots, N$). $x' = x + \text{rand}(0, 1) * |x - xk|$, where x' is the generated point and rand (0, 1) represents the random number between 0 and 1.

The models were implemented in Python language using Jupyter notebook (6.1.4) and sklearn library (0.23.2). For partitioning the data, 10-fold cross-validation technique was used. Experiments were performed on the original dataset and the SMOTE-transformed dataset. Several feature sets were produced using Extratree classifiers with feature importance technique. The set of features was used in the experiments such as all features (25), top 20 features, top 15 features, and top 10 features, respectively. Figure 2 represents the feature ranking, using feature importance, for 20 features.

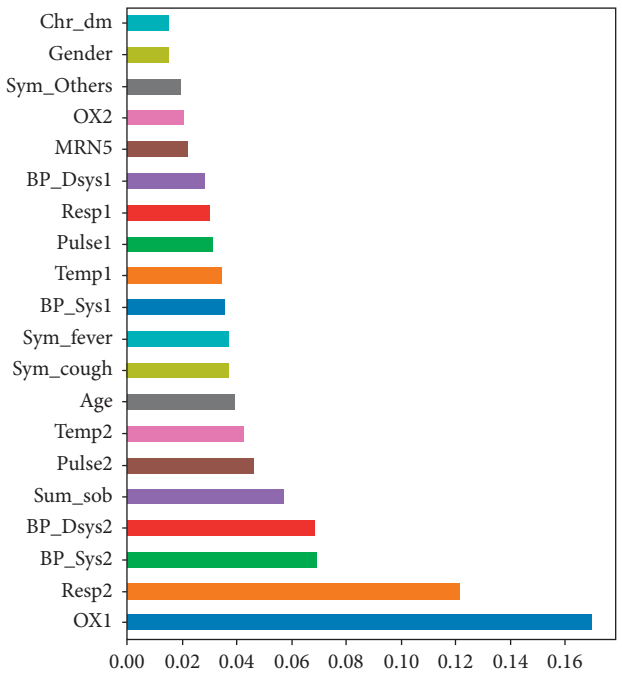

Figure 2: Correlation of top 20 features in the dataset.

The following tables present the performance of the classifiers in terms of accuracy, sensitivity, specificity, and *F*-score. The results showed that random forest outperformed the other models with SMOTE data. Table 7 presents the performance of the classifiers using all features. Table 8 presents the outcome using the top 20 features, Table 9 presents the results with the top 15 features, and Table 10 presents the comparison with the top 10 features, respectively.

Experimental results revealed that random forest outperformed the other classifiers using the top 20 features with SMOTE data with the accuracy of 0.952, sensitivity of 0.949, specificity of 0.956, and *F*-score of 0.955, respectively. Similarly, the AUC-ROC curves for logistic regression, random forest, and extreme gradient boosting are shown in Figures 3, 4, and 5, respectively, using the top 20 features. Random forest achieved the AUC of 0.99. However, the random forest achieved the highest specificity of 1 using the top 15 features.

Logistic regression, on the other hand, underperformed over other classifiers in the top 20, 15, and 10 features using SMOTE data with the accuracy of 0.86, 0.82, and 0.84, respectively. The AUC-ROC curve shows that LR achieved 0.91. However, LR in our study performed better than another study conducted by Yao et al. [13]. They used the LR model to identify the COVID-19 patients' severity and the results achieved an AUC-ROC of 0.881.

A number of studies focused on prediction of severity or mortality have noted that the age is one of the top features that helps to predict the severity of cases [10–13]. In our study, age was ranked among top 10 features across all 25 features used in our prediction model. In addition, our study outperformed other studies that are covered in the literature review with an accuracy of 0.952 and AUC-ROC curve of 0.99.

Table 7: Performance comparison of classifiers using all features (25) using original and SMOTE data.

Classifier	Sampling technique	Accuracy	Sensitivity	Specificity	*F*-score
LR	Without SMOTE	0.874	0.538	0.932	0.56
	With SMOTE	0.753	0.766	0.739	0.766
RF	Without SMOTE	0.908	0.75	0.924	0.6
	With SMOTE	0.938	0.947	0.929	0.941
XGB	Without SMOTE	0.885	0.6	0.922	0.545
	With SMOTE	0.925	0.923	0.926	0.929

Table 8: Performance comparison of classifiers using top 20 features using original and SMOTE data.

Classifier	Sampling technique	Accuracy	Sensitivity	Specificity	*F*-score
LR	Without SMOTE	0.874	0.538	0.932	0.56
	With SMOTE	0.863	0.82	0.93	0.88
RF	Without SMOTE	0.908	0.7	0.935	0.636
	With SMOTE	0.952	0.949	0.956	0.955
XGB	Without SMOTE	0.862	0.5	0.909	0.455
	With SMOTE	0.897	0.878	0.922	0.906

Table 9: Performance comparison of classifiers using top 15 features using original and SMOTE data.

Classifier	Sampling technique	Accuracy	Sensitivity	Specificity	*F*-score
LR	Without SMOTE	0.874	0.583	0.932	0.56
	With SMOTE	0.822	0.793	0.864	0.841
RF	Without SMOTE	0.908	0.7	0.935	0.636
	With SMOTE	0.911	0.856	1	0.922
XGB	Without SMOTE	0.851	0.455	0.908	0.435
	With SMOTE	0.932	0.894	0.984	0.938

This study covers the prediction of the survival and the death of COVID-19-positive patients using demographic, vital signs, and chronic diseases, respectively. The overall result demonstrates the significance of the proposed study with the accuracy of 0.95 and the AUC value of 0.99 using 20 features. The study was performed using a real dataset from the King Fahad University Hospital. Moreover, the dataset

Table 10: Performance comparison of classifiers using top 10 features using original and SMOTE data.

Classifier	Sampling technique	Accuracy	Sensitivity	Specificity	*F*-score
LR	Without SMOTE	0.862	0.5	0.909	0.45
	With SMOTE	0.849	0.867	0.831	0.855
RF	Without SMOTE	0.89	0.63	0.934	0.609
	With SMOTE	0.925	0.884	0.983	0.933
XGB	Without SMOTE	0.851	0.455	0.908	0.43
	With SMOTE	0.89	0.843	0.965	0.904

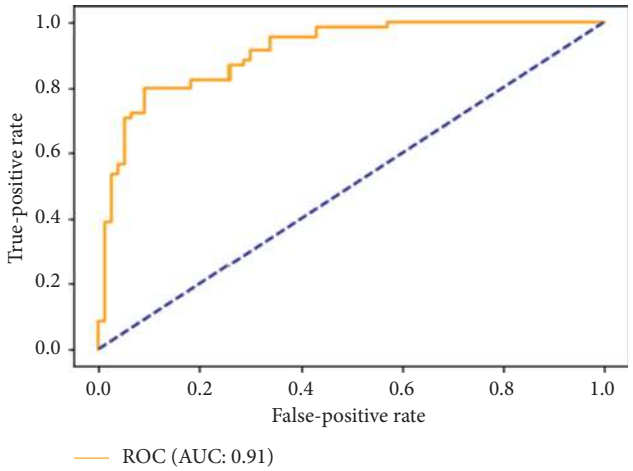

Figure 3: ROC curves of logistic regression using top 20 features.

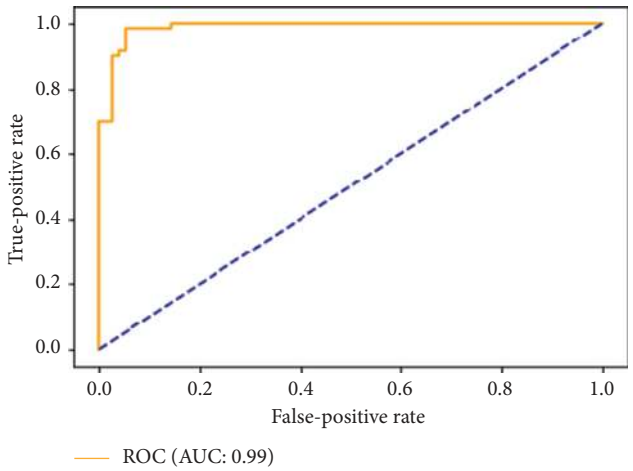

Figure 4: ROC curves of random forest using top 20 features.

contains a very small number of missing data. Despite the several advantages, the study can be further improved by increasing the number of patients. Furthermore, the study needs to incorporate other laboratory tests like lactate dehydrogenase (LDH), neutrophils, lymphocyte, and highly sensitive C-reactive protein. Several identified significant features from the literatures need to be included for predicting the mortality risk in COVID-19 patients.

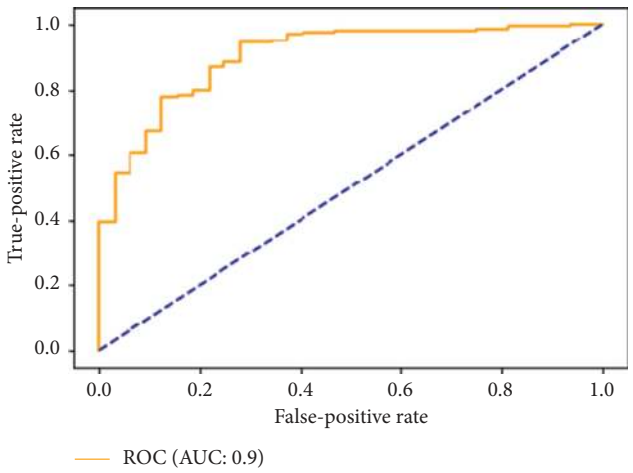

Figure 5: ROC curves of extreme gradient boosting using top 20 features.

4. Conclusion

The COVID-19 pandemic outbreak has devastated the whole world and lead to a state of worldwide health emergency. Several efforts have been performed to combat this pandemic. In this study, we aimed to explore the impact of vital signs, chronic disease, preliminary clinical data, and demographic features to predict the mortality and survival of the COVID-19 patients using supervised machine learning algorithms. Due to the reduced mortality risk of the COVID-19 cases, the dataset suffers from data imbalance. SMOTE technique was used to alleviate the data imbalance. The results showed that random forest outperformed the other models using 10-fold cross-validation. Grid search technique was applied for parameter optimization. The study achieved the accuracy of 0.952 and AUC of 0.99. Despite the significant outcome achieved from this proposed model, there is still a need for improvement. The models need to be validated using multiple datasets. Furthermore, in the future, we will incorporate and explore the impact of other clinical features and laboratory results that were identified as significant in the previous studies.

Data Availability

The data used to support the findings of this study will be shared upon request to the corresponding author, and the IRB details of the data are available in the paper.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi

Arabia for funding this research work through the project number Covid19-2020-059-CSIT at Imam Abdulrahman Bin Faisal University/College of Computer Science and Information Technology.

References

[1] "Worldometers-COVID-19 Coronavirus Pandemic." https://www.worldometers.info/coronavirus/?utm_campaign=homeAdvegas1? (accessed January 17, 2020).

[2] C. Gazzaruso, E. Paolozzi, C. Valenti et al., "Association between antithrombin and mortality in patients with COVID-19. A possible link with obesity," Nutrition, Metabolism and Cardiovascular Diseases, vol. 30, no. 11, pp. 1914–1919.

[3] Z. Malki, E.-S. Atlam, A. E. Hassanien, G. Dagnew, M. A. Elhosseini, and I. Gad, "Association between weather data and COVID-19 pandemic predicting mortality rate: machine learning approaches," *Chaos, Solitons & Fractals*, vol. 138, p. 110137, 2020.

[4] A. S. Albahri, R. A. Hamid, J. K. Alwan et al., "Role of biological data mining and machine learning techniques in detecting and diagnosing the novel Coronavirus (COVID-19): a systematic review," *Journal of Medical Systems*, vol. 44, no. 7, p. 122, 2020.

[5] R. Zagrouba, M. Adnan Khan, A. ur-Rahman et al., "Modelling and simulation of COVID-19 outbreak prediction using supervised machine learning," *Computers, Materials & Continua*, vol. 66, no. 3, pp. 2397–2407, 2021.

[6] A. Kumar, A. Sharma, and A. Arora, "Anxious depression prediction in real-time social data," in *Proceedings of the International Conference on Advances in Engineering Science Management & Technology (ICAESMT)-2019*, pp. 1–7, Dehradun, India, July 2019.

[7] S. Lalmuanawma, J. Hussain, and L. Chhakchhuak, "Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: a review," *Chaos, Solitons & Fractals*, vol. 139, p. 110059, 2020.

[8] M. E. H. Chowdhury, T. Rahman, A. Khandakar et al., "An early warning tool for predicting mortality risk of COVID-19 patients using machine learning," 2020, http://arxiv.org/abs/2007.15559.

[9] M. Nemati, J. Ansary, and N. Nemati, "Machine-learning approaches in COVID-19 survival analysis and discharge-time likelihood prediction using clinical data," *Patterns*, vol. 1, no. 5, p. 100074, 2020.

[10] L. Yan, H.-T. Zhang, Y. Xiao et al., "Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan," *medRxiv*, 2020.

[11] K. C. Y. Wong and H.-C. So, "Uncovering clinical risk factors and prediction of severe COVID-19: a machine learning approach based on UK biobank data," *medRxiv*, 2020.

[12] L. Sun, F. Song, N. Shi et al., "Combination of four clinical indicators predicts the severe/critical symptom of patients infected COVID-19," *Journal of Clinical Virology*, vol. 128, p. 104431, 2020.

[13] H. Yao, N. Zhang, R. Zhang et al., "Severity detection for the coronavirus disease 2019 (COVID-19) patients using a machine learning model based on the blood and urine tests," *Frontiers in Cell and Developmental Biology*, vol. 8, pp. 1–10, 2020.

[14] C. Hu, Z. Liu, Y. Jiang et al., "Early prediction of mortality risk among patients with severe COVID-19, using machine learning," *International Journal of Epidemiology*, vol. 49, no. 6, pp. 1918–1929, 2020.

[15] D. Bertsimas, G. Lukin, L. Mingardi et al., "COVID-19 mortality risk assessment: an international multi-center study," *PLoS One*, vol. 15, no. 12, p. e0243262, 2020.

[16] M. Sánchez-Montañés, P. Rodríguez-Belenguer, A. J. Serrano-López, E. Soria-Olivas, and Y. Alakhdar-Mohmara, "Machine learning for mortality analysis in patients with COVID-19," *International Journal of Environmental Research and Public Health*, vol. 17, no. 22, pp. 8386–20, 2020.

[17] C. An, H. Lim, D.-W. Kim, J. H. Chang, Y. J. Choi, and S. W. Kim, "Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study," *Scientific Reports*, vol. 10, p. 18716, 2020.

[18] R. X. S. D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, John Wiley & Sons, Toronto, Canada, 2013.

[19] Y. M. C. Zhang, *Ensemble Machine Learning*, Springer, New York, NY, USA, 2012.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.