Hindawi
Applied Computational Intelligence and Soft Computing
Volume 2022, Article ID 3783058, 12 pages
https://doi.org/10.1155/2022/3783058

Research Article

Classifying the Mortality of People with Underlying Health Conditions Affected by COVID-19 Using Machine Learning Techniques

Rami Mustafa A. Mohammad,[1] Malak Aljabri,[2,3] Menna Aboulnour,[3] Samiha Mirza,[3] and Ahmad Alshobaiki[3]

[1]*Department of Computer Information Systems, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia*
[2]*Department of Computer Science, College of Computer and Information Systems, Umm Al-Qura University, Makkah 21955, Saudi Arabia*
[3]*Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia*

Correspondence should be addressed to Menna Aboulnour; 2180007190@iau.edu.sa

Received 20 February 2022; Accepted 25 April 2022; Published 17 May 2022

Academic Editor: Manikandan Ramachandran

The COVID-19 pandemic has greatly affected populations worldwide and has posed a significant challenge to medical systems. With the constant increase in the number of severe COVID-19 infections, an essential area of research has been directed towards predicting the mortality rate of these patients, in order to make informed medical decisions about the necessary healthcare priorities. Although a large amount of research has attempted to predict the mortality rate of COVID-19 patients, the association between the mortality rate of COVID-19 patients and their underlying health conditions has been given significantly less attention. Meanwhile, patients with underlying conditions often face a worse COVID-19 prognosis. Therefore, the goal of this study was to classify the mortality rate of patients diagnosed with COVID-19, who also suffer from underlying health conditions or comorbidities. To achieve our goal, we applied machine learning (ML) models on a new publicly available dataset, not investigated by any existing literature. The dataset provides detailed information on 582 COVID-19 patients and facilitates a robust forecasting model of the mortality rate. The dataset was analysed using seven ML classifiers, namely, Bagging, J48, logistic regression (LR), random forest (RF), support vector machine (SVM), naïve Bayes (NB), and threshold selector. A comparative analysis was performed across the seven ML techniques, and their performance was assessed based on evaluation parameters including classification accuracy, true-positive rate, and false-positive rate. The best performance was demonstrated by the Bagging algorithm with an accuracy of 83.55% when using all the dataset features. The findings are intended to assist researchers and physicians in the early identification of at-risk COVID-19 patients and to make the appropriate intensive care decisions.

1. Introduction

Being a large family of RNA viruses known to have existed since the mid-1960s, coronaviruses usually cause mild to moderate upper-respiratory tract illnesses such as the common cold [1, 2]. However, in the past decade, several new coronaviruses have mutated and caused serious illness, globally. Three of the most serious coronaviruses are known to be severe acute respiratory syndrome-coronavirus (SARS-CoV), Middle East respiratory syndrome coronavirus (MERS-CoV), and severe acute respiratory syndrome-coronavirus 2(SARS-CoV-2) causing COVID-19, which emerged lately from the Chinese city of Wuhan in December 2019 [3]. COVID-19 is a contagious disease that causes respiratory diseases of varying severity ranging from regular flu symptoms to serious illnesses and could even lead to

death. From December 2019 to February 2020, the world witnessed such a massive spread of COVID-19 infections leading the World Health Organization (WHO) to declare it as a global pandemic [3]. As of February 2022, more than 420 million cases have been reported worldwide, with a mortality rate of around 1.4% of the reported cases [4]. The pandemic has greatly impacted the lives of people around the world, in many ways, and has especially challenged the health system and governments globally. Most people diagnosed with COVID-19 suffer mild to moderate symptoms and regain their health without necessitating special treatment. However, some COVID-19 patients become extremely ill and require specific medical attention. Therefore, managing the number of COVID-19 cases has been a huge challenge for healthcare facilities globally.

Due to the noticeable dangerous/serious effects caused by the virus, immediate research efforts needed to be carried out to gain a better understanding of the situation and provide the best solutions regarding the issues faced by society due to the spread of COVID-19. An essential characteristic of any crucial disease, especially one caused by dangerous coronavirus mutants, is the measure of its ability to lead to eventual death. Thus, predicting mortality rates aids scientists and physicians to understand the severity of a disease such as COVID-19, identifying the populations at risk, and evaluating the quality of necessary health care [5]. Furthermore, differentiating the patients with severe or nonsevere COVID-19 infections is beneficial for a timely decision on the clinical monitoring level required. For instance, patients with a low COVID-19 mortality rate can be accommodated with less intensive clinical monitoring. Patients with a high rate, on the other hand, must be admitted to an intensive care unit (ICU) by clinicians, as they require constant monitoring [6]. Hence, determining the severity or mortality rates of the affected COVID-19 patients is essential. Therefore, to address these needs, existing clinical and medical laboratory tests are being used to determine the patient's mortality risk. However, these techniques are often time-consuming and require years of medical experience [7]. Alternatively, ML techniques are actively being explored to understand and combat COVID-19, due to their ability to extract essential knowledge from collected data and, thus, aid in decision making. Considering their success, ML models have gradually become a reliable aid in numerous healthcare services.

Current research analysis has forwarded concern that people with underlying health conditions have a worse COVID-19 prognosis [8, 9]. These patients have been identified as particularly vulnerable to greater morbidity and mortality risks when diagnosed with COVID-19. Several medical studies [8, 10, 11] affirm that comorbidities are an indication of higher death rates among COVID-19 patients. Most of these studies employed a medical research strategy that involves clinical and medical tests in their investigations. However, our objective in this study is to employ ML techniques to classify the mortality risk of patients who have underlying health conditions or comorbidities and have also been diagnosed with COVID-19. To achieve our goals, we developed seven ML models, namely, Bagging, J48, LR, NB, RF, SVM, and Threshold Selector. Our aim is to build acceptable classifiers trained on a new dataset of COVID-19 patients, to classify the mortality rate of patients diagnosed with COVID-19 that also suffer from underlying health conditions and, thus, gain better insights on the issue.

The main contributions of this paper are as follows:

(1) Demonstrating the significance of the association between the underlying health conditions and the prognosis of COVID-19 cases using ML techniques.

(2) Studying the correlation between the different features, representing underlying health conditions in the dataset, and maintaining only the most relevant ones based on different feature selection techniques to show the significance of each of the features on patient mortality.

(3) Conducting a comparative performance evaluation by applying seven ML models to identify the mortality risk of COVID-19 patients with underlying health conditions.

This paper is structured as follows: Section 2 presents a review of the existing literature on the related topic. Section 3 describes the materials and methods used to conduct the study. Section 4 discusses the experimental setup and the results. Finally, Section 5 concludes the paper.

2. Literature Review

A few studies have been carried out focusing on identifying the association between underlying health conditions and the mortality of COVID-19 patients using ML models. On a more general approach, some studies worked on introducing ML models to support the prediction of the mortality rate of hospitalized COVID-19 patients regardless of their underlying health conditions. In this section, we will review the existing literature on the topic, in order to assess and identify the gaps and add valuable contributions.

Several studies used ML to investigate the course and progression of COVID-19 infections, in patients who suffer from underlying health conditions. Following this idea, García-Azorín et al. [10] analysed whether the presence of chronic neurological disorders (CND) in COVID-19 patients is an indicator of elevated mortality risk. Patients' survival time was analysed using a cox regression log-rank test on 576 patients diagnosed with COVID-19. The results showed that the presence of CND is an objective predictor of death, with a confidence interval of 95%. However, this study did not investigate the mortality risk of CND patients suffering from severe cases of COVID-19. In another study, Roy et al. [12] investigated the mortality of COVID-19 patients with inflammatory bowel disease (IBD), on 20,000 IBD patients. SVM, stochastic gradient decent (SGD), nearest centroid (NC), DT, GNB, and MLP were applied using cross-fold validation to determine primary and secondary covariates to predict the mortality of the patients. The analysis revealed that primary covariates are age, medication usage, and the number of comorbidities, while the secondary features were IBD severity, smoking history, gender, and IBD subtype. Similarly, Pérez et al. [11]

performed an evaluation of factors, such as clinical features, prognostic factors, and comorbidities related to in-hospital mortality of 96 COVID-19 patients. The study found that the most recurrent comorbidities were hypertension in 40% of the patients, diabetes mellitus in 16% of the patients, and cardiopathy in 14% of the patients. Through their analysis, they concluded that the variables with the highest association with the risk of death during a hospital stay were the presence of cardiopathy, an increase of lactate dehydrogenase (LDH) levels to more than 345 IU/L, and an age of more than 65 years.

Furthermore, Sanyaolu et al. [8] also investigated the progression, comorbid conditions, and mortality rates of 1,786 patients diagnosed with COVID-19. They identified that the most common conditions were hypertension, present in 15.8% of the patients; cardiovascular and cerebrovascular conditions, present in 11.7%; and diabetes, present in 9.4%. They concluded that patients with comorbidities experience a more severe prognosis. Moreover, patients who have a record of hypertension, chronic lung disease, obesity, diabetes, and cardiovascular disease have the worst prognosis and are usually associated with more severe outcomes such as acute respiratory distress syndrome (ARDS) and pneumonia. Likewise, Kang [13] found that patients suffering from at least one underlying condition demonstrated a higher death rate, especially in patients that were older than 70. The most common underlying conditions were diseases in the circulatory system, such as arrhythmia, myocardial infarction, cerebral infarction, and hypertension. Moreover, Banerjee et al. [14] aimed to gain more knowledge about the high mortality of the COVID-19 pandemic, based on sex, age, and underlying conditions, by assessing 3 million individuals. They found that 68.5% of the patients in the high-risk category were older than 70 years and the remaining 31.2% suffered from at least one underlying disease. Hence, they concluded that age and underlying conditions significantly impact the level of risk on the COVID-19 patient. Additionally, Kompaniyets et al. [9] evaluated the risk of a critical COVID-19 prognosis across children and its association with underlying conditions. They conducted a cross-sectional study that included 43,465 children diagnosed with COVID-19 and used generalized multivariate linear models. The most observed conditions were asthma and neurodevelopmental disorders. However, the most significant indicators of hospitalization were type 1 diabetes and obesity. Furthermore, the most significant indicators for the critical prognosis of COVID-19 were cardiac and circulatory abnormalities and type 1 diabetes.

Some studies focused on predicting the mortality rate of hospitalized COVID-19 patients. Following this principle, Guan et al. [15] used the least absolute shrinkage and selection operator (LASSO) method to screen 48 clinical and laboratory features on a dataset of 1,270 hospitalized patients and applied an extreme gradient boosting (XGBoost) method to predict the death risk. Six features, namely, severity, age, levels of high-sensitivity C-reactive protein, lactate dehydrogenase, ferritin, and interleukin-10, were selected, and the method obtained a precision of 90%. Similarly, Tezza et al. [16] aimed to identify the indicators of COVID-19 in-hospital mortality by comparing the performance of multiple ML algorithms such as recursive partition tree (RPART), gradient boosting machine (GBM), SVM, and RF. A dataset of 341 patients was used to train the models. The RF algorithm achieved the highest performance with a receiving operative characteristic (ROC) of 0.84. They concluded that the strongest indicators of in-hospital mortality were age, along with vital signs, and laboratory results. Furthermore, Parchure et al. [17] built a model for predicting mortality of in-hospital COVID-19 patients on a dataset of 567 patients. The RF classifier was used, and the input features included patients' laboratory results, electrocardiogram (ECG) results, and vital signs. The model achieved an area under the curve (AUC) of 85.5%. On the other hand, Subudhi et al. [18] used ML algorithms such as RF, NB, logistic regression, and ensemble models to predict ICU admission and mortality using a dataset of nearly 5,000 COVID-19 patients. The results found that ensemble models were better at predicting the mortality rates. Furthermore, features such as oxygen saturation and glomerular filtration rate were useful in determining the likelihood of admission to the ICU. Similarly, Chowdhry et al. [19] created an ML prediction model using XGBoost for early warning of mortality risk using a dataset from a study conducted by Yan et al. [20]. Features acquired at hospital admissions, namely, lactate dehydrogenase, neutrophils, lymphocyte, high-sensitivity C-reactive protein, and age, were identified as key predictors of death by the model. The model obtained an AUC of 99.1%. In another study, Gao et al. [21] focused on creating an early warning system using an ensemble of LR, SVM, gradient boosted decision tree, and neural networks. The results reached an AUC of 96.2%. Likewise, Pourhomayoun and Shakibi [22] applied SVM, RF, DT, LR, K-nearest neighbour (KNN), and artificial neural network (ANN) on a dataset containing two million COVID-19 patients' records to support medical decisions and determine health risk. The overall accuracy of predicting the mortality rate was 89.9%.

Many studies worked on developing ML models to automate the overall prediction of COVID-19 patients' mortality. For instance, Aljameel et al. [23] proposed a method for the early prediction of the outcomes of COVID-19 patients by comparing three classification techniques, namely, LR, XGBoost, and RF. The models were built using 287 COVID-19 patients and 20 clinical features. The RF classifier achieved the best performance with an AUC of 99%. Additionally, Khan et al. [24] used ML algorithms such as DT, LR, KNN, XGBoost, RF, and deep learning (DL) model with six layers to forecast the mortality rate in patients diagnosed with COVID-19. The models were developed using 103,888 patient records and a comparative analysis was conducted where the best performance accuracy of 97% was obtained using the DL model. Similarly, Booth et al. [25] also developed an SVM model to predict COVID-19 mortality in 398 patients and obtained an AUC of 93%. Similarly, many other studies [26–32] focused on the same ideology of predicting mortality rates using ML models and obtained results of various degrees of accuracy.

Overall, most of the studies discussed above either prove the effect of certain underlying health issues and the progression of COVID-19 cases or demonstrate the importance of ML in forecasting the mortality rate of patients suffering from COVID-19. However, much of the existing research that introduced ML models for COVID-19 mortality prediction did not consider the association between the patient's mortality rate and their underlying health conditions. Hence, this research is essential, since people with underlying health conditions affected by COVID-19 have a worse prognosis. Moreover, early mortality risk prediction can aid physicians in deciding the necessary actions and treatment, such as admitting the patient early into the ICU. Thus, the main objective of our study is to highlight the connection between the patients' mortality rate and their underlying health conditions and draw healthcare benefits from the results. Furthermore, we employed a new dataset released by the Harvard Dataverse [33] that has not been used in any studies previously. Our study thus accommodates to the changing nature of the effects of COVID-19 on different patients and demonstrates the importance of experimenting with new data for relevant discovery. Finally, we performed multiple experiments with different feature selection techniques to enhance the performance of the classifiers and to identify the features responsible for making these performance enhancements possible.

3. Materials and Methods

3.1. The Dataset. The "Replication Data for: Ethnicity, pre-existing comorbidities, and outcomes of hospitalized patients with COVID-19" dataset is available at Harvard Dataverse [33]. It contains the health condition and attributes that contribute to the outcomes of patients with COVID-19. The dataset contains the demographic, ethnic, socioeconomic, and clinical risk factors associated with the outcomes of COVID-19 patients. Furthermore, the dataset consists of 582 detailed instances of COVID-19 inpatients of which 408 are White, 142 are South Asian, and 32 are other minority ethnic patients. Severe risk factors have been identified as sex, age, obesity, and pre-existing comorbidities such as hypertension, diabetes, coronary heart disease, chronic obstructive pulmonary disease, asthma, chronic renal disease, and cancer. The dataset has 17 attributes as shown in Table 1.

3.2. Methodology. The main purpose of our study is to use ML techniques to classify the mortality rates of patients suffering from COVID-19 while taking into account their underlying health conditions. The models used include Bagging, J48, LR, NB, RF, SVM, and Threshold Selector. The models were trained using the previously mentioned dataset with the aim of predicting the expected mortality of the patients. Furthermore, we evaluated the performance of these models based on evaluation parameters including classification accuracy, true-positive rate (TPR), and false-positive rate (FPR). We used the 70-30% holdout split to build the models and performed three experiments using different subsets of features to test the significance of the features and increase the performance. Figure 1 shows the overall structure of the research methodology.

3.3. Preprocessing. Preprocessing is performed before any analysis, to ensure that the data are suitable for training and testing the models. This process includes loading, cleaning, handling, and transferring the data into a proper format for the intended tasks. The "ID" attribute was removed as it does not provide any real insight during the classification process. The dataset contained 582 instances of which only 189 are patients that passed away within 30 days. These patients are marked as "yes" in the dataset. Patients that passed away represent 32% of the dataset, which indicates that the dataset suffers from imbalance, where one class is noticeably less represented compared to the other. Since the dataset is imbalanced and the number of instances is limited, we applied oversampling on the dataset to make the number of instances per category nearly equal. We applied the Synthetic Minority Oversampling Technique (SMOTE) filter to double the number of "yes" instances by randomly duplicating its instances as demonstrated in Figure 2 and Table 2. Then, the randomize filter was applied to avoid overfitting. Finally, label encoding was used to convert categorical features into numerical values.

3.4. Feature Selection. The dataset included 17 features describing relevant information about the patients. We conducted three different experiments to reach the optimal performance and contributed to improving the past efforts achieved by previous researchers. The "Death30 days" attribute was selected as the class attribute, which takes the values of "yes," or "no." The first experiment included all the features with the exception of the "ID" attribute due to its irrelevance in model training. For the second experiment, the "Attribute Selection" supervised attribute filter with the "CfsSubsetEval" evaluator and the "BestFirst" search method were applied. For the third experiment, we eliminated 4 features, the "Renal disease," "Cancer," "Copd," and "ICU," based on their low correlation with the class attribute as shown in Table 3. Despite "Diabetes1" displaying a low correlation, it was not eliminated as it highly affects the performance of several models. Table 4 shows the different sets of features used in each of the three experiments.

3.5. Proposed Techniques

3.5.1. Bagging. Bagging is an ensemble model that aims to increase the accuracy and stability of ML algorithms used in statistical classification and regression. In addition, it avoids overfitting by reducing variance. The bagging process selects an arbitrary sample of data from the training set with replacement, indicating that each data point could be selected more than once. Due to significant variation or bias, a single model, also known as a base or weak learner, may not perform effectively on its own. Therefore, these weak models are trained separately, and the performance average of those weak learners

Table 1: Description of dataset attributes.

Attribute name	Type	Description
ID	Numeric	Row id
Agecat	Numeric	Patient's age group
Sex	Nominal	Patient's gender
Ethnicity3cat	Nominal	Patient's ethnicity
Imd	Nominal	Indices of multiple deprivation (A measure of poverty)
Bmicat	Nominal	Patient's BMI
Diabetes1	Nominal	Patient's diabetes type 1 status—yes/no
Diabetes2	Nominal	Patient's diabetes type 2 status—yes/no
Hypertension	Nominal	Hypertension status—yes/no
Cvd	Nominal	Chronic heart disease—yes/no
Asthma	Nominal	Asthma status—yes/no
Copd	Nominal	Chronic obstructive pulmonary disease—yes/no
Cancer	Nominal	Cancer status—yes/no
Renal disease	Nominal	Renal disease status—yes/no
ICU	Nominal	Whether the patient was admitted into the ICU—yes/no
Death30 days	Nominal	If patient died within 30 days of infection
Timeatrisk	Nominal	Duration of infection

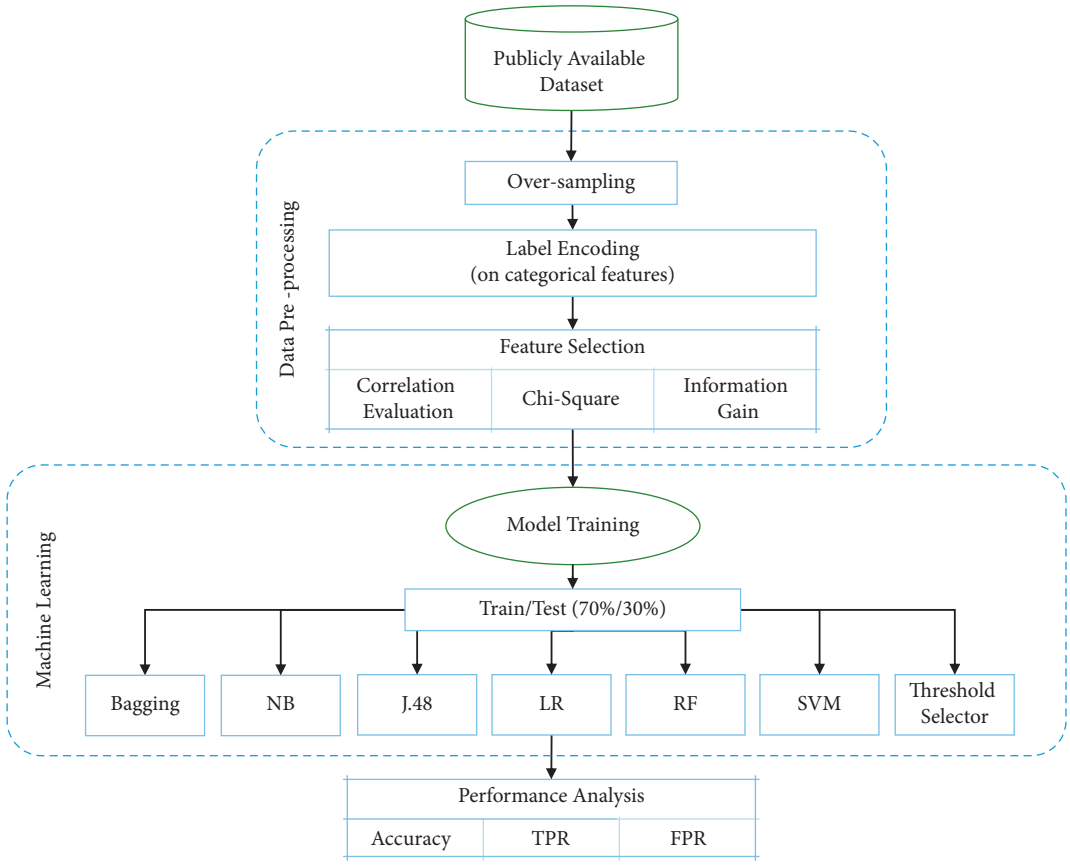


Figure 1: Research methodology steps.

is combined to produce a strong learner, an ensemble classifier. Hence, their predictions yield more accurate results [34].

3.5.2. Logistic Regression. LR is a ML model that is applied to solve classification problems using a predictive analytic approach based on the notion of probability. The LR classifier employs a complicated cost function, which is referred to as the "Sigmoid function" or the "Logistic function." The LR hypothesis limits the sigmoid function to a value between 0 and 1. In ML, the sigmoid function is used to map the predictions to probabilities [35].

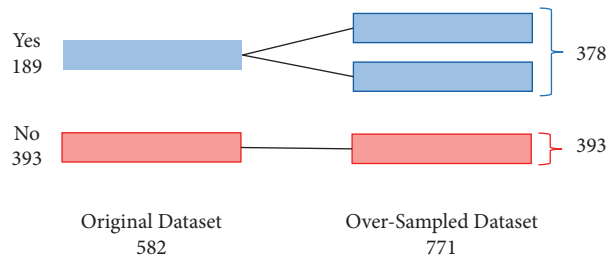


Figure 2: Dataset statistics before and after applying randomized oversampling.

Table 2: Dataset values before and after oversampling.

Death30 days	Before	After
Yes	189	378
No	393	393
Total	582	771

3.5.3. Random Forest. RF is an ensemble algorithm that generates an estimate of the expected result by combining many different decision trees and calculating the average of all their prediction results [33]. Thus, the RF algorithm is an extension of the Bagging technique.

3.5.4. Support Vector Machine. The SVM classifies data based on class attribute value and produces the best possible hyperplane. The hyperplane is a line in two dimensions that serve as a decision boundary to optimally separate the predictions. Thus, everything falling on one side of the hyperplane will be categorized as belonging to one class, while anything falling on the other side will be categorized into another class [36]. Hence, the working premise of the SVM is to draw a line that divides the data into two categories and distinguishes between them.

Table 3: Correlation evaluation of the dataset features.

Feature	Correlation
Cvd	0.25807
Agecat	0.20561
Ethnicity3cat	0.18621
Sex	0.17204
Asthma	0.16892
Hypertension	0.12359
Timeatrisk	0.12305
Diabetes2	0.1034
Imd	0.09377
Bmicat	0.08311
Renal disease	0.07616
Diabetes1	0.04363
Cancer	0.03207
Copd	0.00985
ICU	0.00429

3.5.5. Naïve Bayes. The NB algorithm is based on the concept of the conditional probability of the Bayes theorem formulated by Thomas Bayes. The probability that an event will occur if another event has already taken place is known as conditional probability. We can calculate the likelihood of the occurrence of an event by using past knowledge and the conditional probability, as depicted by

$$P(A|B) = \frac{P(B \mid A)P(A)}{P(B)}. \tag{1}$$

The Bayes theorem is used as a fundamental theorem for the NB classifier. It calculates the association probabilities for each class, such as the likelihood that a certain data point or instance belongs to a specific class, as shown in equation (2). It considers each attribute independently, so a feature's presence or absence has no relevance to the presence or absence of any other features [37].

$$\frac{P(H|\text{Multiple evidences}) = P(E1|H)^{*}P(E2|H)||^{*}\ldots P(En|H)^{*}P(H)}{P(\text{Multiple evidences})}. \tag{2}$$

3.5.6. J48. J48 is an ML decision tree (DT) algorithm. Generally, a DT algorithm has a root node, intermediate nodes, and leaf nodes. Furthermore, each node in the tree represents a decision that leads from the root to a leaf node representing the final result. The input data are divided into mutually exclusive regions by an attribute, and each region represents a value, label, or action to characterize its data points. The dividing criterion determines which attribute is optimal to be used to split that tree section [37].

3.5.7. Threshold Selector. The Threshold Selector is a meta-classifier that works on choosing a midpoint threshold on the results output by another algorithm. Setting a midpoint threshold aims to optimize the performance of the algorithm used. It is beneficial to apply the Threshold Selector when the algorithm produces results that are within a tight range, as the Threshold Selector expands the range of the algorithm's results to improve its performance [38].

4. Evaluation Metrics

This section presents the evaluation parameters used to assess the models' performance. In this paper, we assessed the classification performance of the models based on classification accuracy, TPR, and FPR. The classification accuracy represents the ratio of successfully predicted instances and is calculated using (3) by finding the ratio of the number of correct predictions to the total number of predictions. The true positive (TP) and true negative (TN) refer

Table 4: Features used in each experiment.

	Experiment 1: All dataset features included	Experiment 2: After applying the "attribute selection" filter	Experiment 3: Eliminating the features with the lowest correlation
Features	Agecat	Agecat	Agecat
	Cvd	Cvd	Cvd
	Ethnicity3cat	Sex	Ethnicity3cat
	Sex	Asthma	Sex
	Asthma	Timeatrisk	Asthma
	Hypertension	Imd	Hypertension
	Timeatrisk		Timeatrisk
	Diabetes2		Diabetes2
	Imd		Imd
	Bmicat		Bmicat
	Renal disease		Diabetes1
	Diabetes1		
	Cancer		
	Copd		
	ICU		

to instances that were correctly classified to positive and negative labels, respectively. Meanwhile, false positive (FP) and false negative (FN) represent incorrectly classified instances to positive and negative labels, respectively.

$$\text{Classification accuracy} = \frac{TP + TN}{TP + FN + FP + TN'}. \quad (3)$$

In addition, TPR represents the ratio of patients who have passed away and were correctly predicted as "yes." In other words, it evaluates how effective the model is at predicting the probability of a patient's death. The TPR can be calculated using equation (4). The higher the TPR of the model, the lower the false-negative rate (FNR) becomes.

$$\text{TPR} = \frac{TP}{TP + FN}. \quad (4)$$

As for FPR, it is the ratio of incorrect predictions of patients that have not passed away yet were incorrectly predicted as "yes." In other words, it evaluates how likely the model is to make incorrect predictions regarding the probability of the patient's death. The FPR can be calculated using thefollowing equation [37]:

$$\text{FPR} = \frac{FP}{FP + TN}. \quad (5)$$

5. Description of the Experiments

In the ML field, supervised ML algorithms have been a popular strategy, especially when dealing with health data due to their ability to learn from the labelled data and effectively predict the disease in question [39]. The goal of this research is to discover key trends in patients with underlying health conditions diagnosed with COVID-19 using several supervised ML algorithms. In a study published in the BMC Journal of Medical Informatics and Decision Making, substantial investigation was carried out to find medical research articles that used more than one supervised ML algorithm to predict a particular disease. This research gives a comprehensive assessment of the relative performance of various supervised ML algorithms for disease prediction. This crucial knowledge about the relative performance of the algorithms can be beneficial in assisting researchers in choosing the best supervised ML algorithm to implement in their research [40].

Their results led them to conclude that the SVM algorithm was found to be the most widely used, followed by the NB algorithm. However, RF showed greater accuracy when applied in some other studies. In such studies, SVM usually exhibited the second highest accuracy. Therefore, in accordance with the findings of the mentioned research, we proposed to use SVM, NB, RF, and J48 algorithms in addition to Bagging, LR, and Threshold Selector. Furthermore, the dataset used is as mentioned before, the "Replication Data for: Ethnicity, pre-existing comorbidities, and outcomes of hospitalized patients with COVID-19," which contains 771 instances after oversampling was applied. Three experiments were carried out on the dataset; the first included 15 features, the second included six features, and the third included 11 features. Moreover, the data were split into two sections of 70-30% for training and testing, respectively. Initially, we conducted the first experiment using all seven models to obtain a baseline accuracy. Subsequently, the second and third experiments were conducted, and all changes to parameter setting and performance measures were reported accordingly. All these details are demonstrated in the following sections.

6. Parameter Optimization

The cross-validation (CV) parameter selection algorithm was used to tune the parameters of all the ML models applied. The final parameter tuning settings for each of the models are displayed in Table 5.

7. Results and Discussion

The first experiment was conducted using 15 features, the second experiment was conducted using 6 features, and the

Table 5: Parameter optimization settings.

Model	Parameter	Optimal value		
		Exp. 1	Exp. 2	Exp. 3
SVM	Kernel	Linear	RBF	RBF
	Cost	20	1	1
RF	Iterations	962	68	120
	Batch size	100	100	100
	Features	0	0	0
J48	Binary split	False	True	False
	Confidence factor	0.25	0.15	0.25
NB	Kernel estimator	False	True	True
	Supervised discretization	True	False	False
LR	Maxlts	−1	−1	−1
	Ridge	1.0E-8	1.0E-8	1.0E-8
Bagging	Classifier	REPTree		
	Iterations	20	32	49
	bagSizePercent	100	100	100
Threshold selector	Classifier	Logistic		
	Designated class	Class value name	First class value	
	Measure	F-measure	Accuracy	

third experiment was conducted using 11 features as mentioned in Section 3.4, and the class label is "Death30 days," which takes either "yes" or "no." The ML models used are SVM, RF, J48, NB, LR, Bagging, and Threshold Selector. The classification accuracy, TPR, and FPR are used to evaluate the performance of the models in all experiments, and the results are demonstrated in Table 6 and Figure 3.

As shown in Table 6, Bagging obtained the best results in terms of all the evaluation matrices in the first and third experiments with accuracies of 83.55% and 83.117%, respectively. For the second experiment, the LR algorithm achieved the best performance accuracy of 81.818% and the best TPR and FPR of 0.818, and 0.175, respectively. However, it is evident that all the models achieved relatively good results with respect to all the evaluation matrices. As for the variation in performance across the experiments, although the difference is generally small, most of the models performed better in the first experiment.

Table 7 lists the best accuracy, TPR, and FPR values obtained by each of the seven models used across the three experiments in descending order in terms of accuracy.

As demonstrated in Table 7, Bagging presented the highest accuracy in the first experiment, which entailed using all the dataset features. As mentioned in Section 3.5, being an ensemble model, Bagging increases the accuracy of prediction by training multiple REPTrees separately and combining the average or most of these weak learners in turn producing a strong ensemble classifier. Using all features of the dataset in the first experiment, after running the Bagging algorithm on a weak learner, REPTree, it produced an accuracy of 83.55% with a TPR of 0.835 and a FPR of 0.160. The classifier correctly classified 91 TP instances and incorrectly classified 10 FP instances. In the sampled set, the model correctly predicted the mortality outcome of 193 patients, correctly predicting 91 patients would survive, and 102 would die within 30 days of infection. However, the model incorrectly predicted the mortality outcome of 38 patients, falsely predicting that 10 patients would survive, and 28 patients would die within 30 days of infection, yet they did not. Performance analysis of the Bagging classifier is shown in Table 8.

Table 6: Experiment results.

Model	Evaluation matrix	Exp. 1	Exp. 2	Exp. 3
SVM	Accuracy	81.385%	78.355%	77.922%
	TPR	0.814	0.784	0.779
	FPR	0.181	0214	0.220
RF	Accuracy	82.251%	74.026%	80.52%
	TPR	0.823	0.740	0.805
	FPR	0.175	0.258	0.193
J.48	Accuracy	78.355%	78.788%	76.623%
	TPR	0.784	0.788	0.766
	FPR	0.215	0.209	0.232
NB	Accuracy	78.788%	80.087%	79.654%
	TPR	0.788	0.801	0.797
	FPR	0.207	0.194	0.198
LR	Accuracy	82.251%	**81.818%**	79.221%
	TPR	0.823	0.818	0.792
	FPR	0.173	0.175	0.202
Bagging	Accuracy	**83.55%**	78.788%	**83.117%**
	TPR	0.835	0.788	0.831
	FPR	0.160	0.207	0.165
Threshold selector	Accuracy	80.952%	81.385%	80.087%
	TPR	0.810	0.814	0.801
	FPR	0.185	0.181	0.200

The LR classifier produced the second highest overall accuracy when run in the second experiment. In the first experiment, after running the LR algorithm, it produced an accuracy of 82.251% with a TPR of 0.823 and an FPR of 0.173, which is only slightly less than the results produced by the Bagging algorithm in the same experiment. In the second experiment, LR correctly classified 90 TP instances and

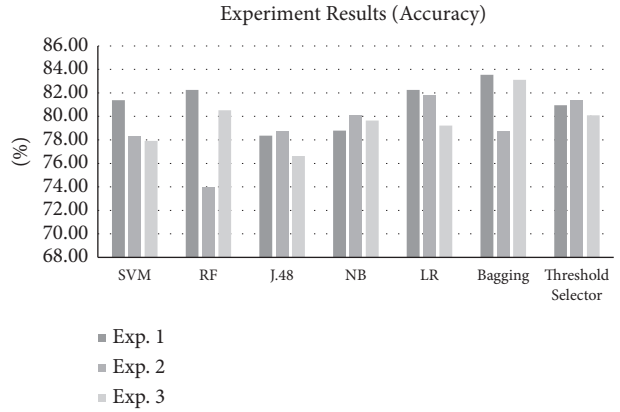
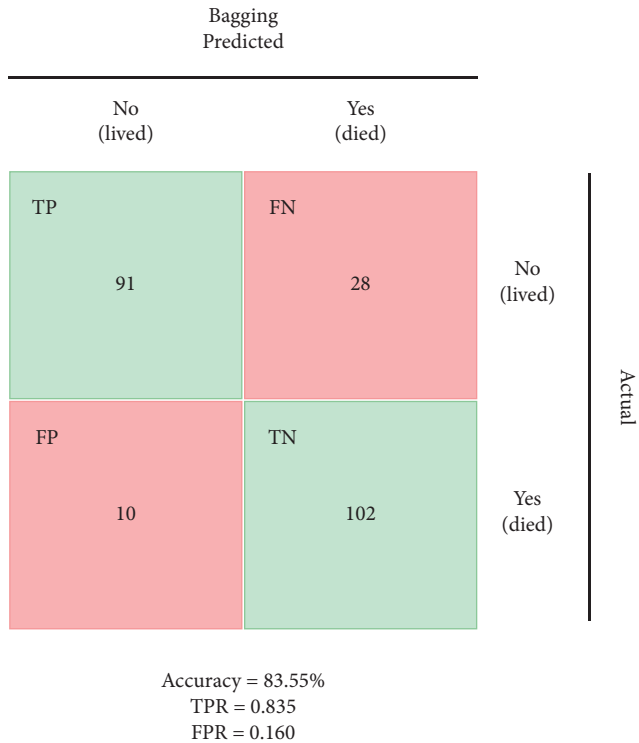


Figure 3: Comparing the accuracies achieved by each of the classifiers in all three experiments.

Table 7: Best performance of each model.

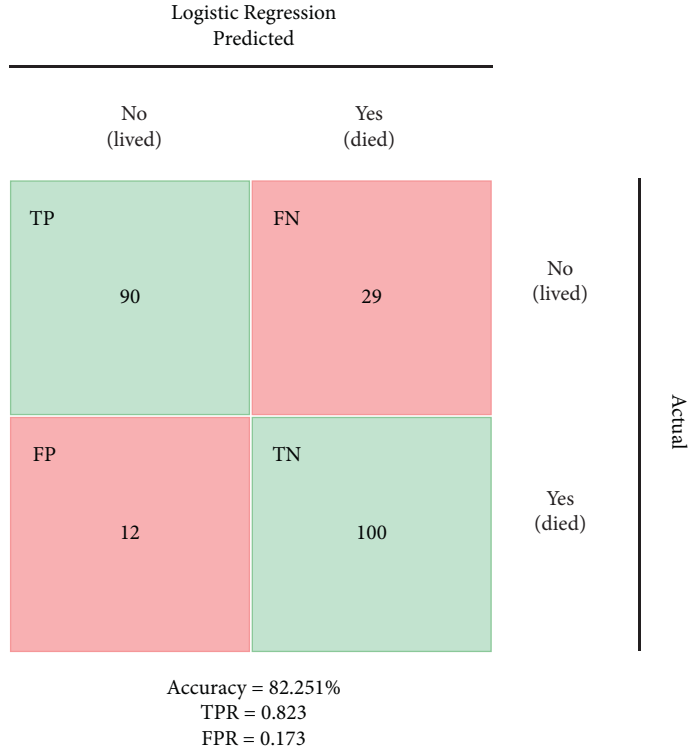
Model	Exp. Num	Accuracy (%)	TP rate	FP rate
Bagging	1	83.55	0.835	0.160
LR	1	82.251	0.823	0.173
RF	1	82.251	0.823	0.175
SVM	1	81.385	0.814	0.181
Threshold selector	2	81.385	0.814	0.181
NB	2	80.087	0.801	0.194
J.48	2	78.788	0.788	0.209

Table 8: Bagging performance analysis.



incorrectly classified 12 FP instances. In the sampled set, the model correctly predicted the mortality outcome of 190 patients, correctly predicting 90 patients would survive, and 100 would die within 30 days of infection. However, the model incorrectly predicted the mortality outcome of 41 patients, falsely predicting that 12 patients would survive, and 29

Table 9: LR performance analysis.



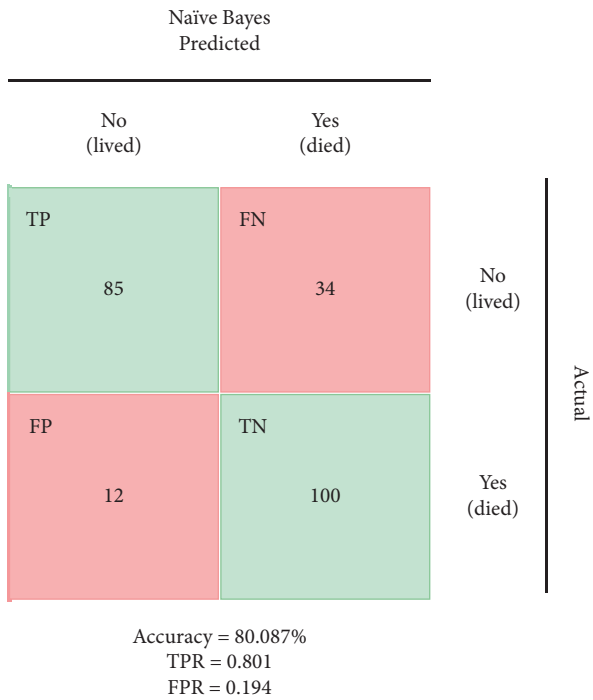
patients would die within 30 days of infection, yet they did not. Table 9 shows the performance analysis of the LR classifier.

The RF algorithm presented the same accuracy as the LR algorithm in the first experiment. Additionally, both RF and LR have the same performance measures, except that RF has a slightly higher FPR. Nevertheless, the difference is negligible. The SVM algorithm produced an accuracy of 81.385% in the first experiment. In this experiment, after running the SVM with a linear kernel type and cost parameters of 20, it produced an accuracy of 81.385% with a TPR of 0.814 and an FPR of 0.181. In addition, the Threshold Selector algorithm in the second experiment using the LR classifier with the F-measure parameter produced the same results as the SVM.

The NB classifier presented an accuracy of 80.087% in the second experiment. It is known that NB considers each attribute independently, so a certain feature's presence or absence has no relevance to the presence or absence of any other feature, an assumption making the classifier simple. However, due to this assumption, its performance is negatively affected when there are redundant or highly correlated features [41]. Therefore, we applied an Attribute Selection filter with "CfsSubsetEval" evaluator and the "BestFirst" search method. The performance analysis of the NB classifier is shown in Table 10.

The NB classifier presented its best accuracy of 80.087% with a TPR of 0.801 and an FPR of 0.194 in the second experiment. In the second experiment, certain features were selected based on their correlation with the target class. The confusion matrix in Table 10 shows that the classifier correctly classified 85 TP instances and incorrectly classified 12 FP instances. In the sampled set, NB correctly predicted the mortality outcome of 185 patients, correctly predicting 85 patients would survive and 100 patients would die within 30 days after infection. However, the model incorrectly predicted the mortality outcome of 46 patients, falsely predicting 12 patients would survive and 34 patients would die within 30 days of infection, yet they did not.

Table 10: NB performance measure.



Lastly, the J48 classifier presented an accuracy of 78.788% in the second experiment. To get the most out of the J48 model, we also applied the Attribute Selection filter to remove redundant attributes. In the second experiment, after running the algorithm, J48 produced an accuracy of 78.788%, a TPR of 0.788, and a FPR of 0.209.

8. Conclusion

Due to severe coronavirus mutations, the COVID-19 pandemic emerged and negatively affected the lives of people around the world, especially challenging the healthcare systems. With the rise in the number of severe COVID-19 cases globally, researchers have directed their efforts towards measuring the likelihood of COVID-19 infections leading to the eventual death of patients. Predicting mortality rates of COVID-19 patients was found to significantly aid scientists and physicians in understanding its severity, level of risk, and most importantly, evaluating the quality of health care needed by any respective patient. Moreover, research has also proven that patients suffering from underlying health conditions face worse prognoses. Hence, the association between the mortality rate of COVID-19 patients and their underlying health conditions was an important topic to discuss. However, there was a lack of research regarding this issue. Therefore, in this study, we focused on classifying the mortality outcome of people suffering from underlying health disorders or comorbidities, who have been diagnosed with COVID-19, in order to aid clinicians and physicians in deciding the appropriate medical attention necessary. To develop a novel solution, we used a ML approach where we employed a recent dataset to classify the mortality of people suffering from COVID-19 and underlying illnesses. The "Replication Data for: Ethnicity, pre-existing comorbidities, and outcomes of hospitalized patients with COVID-19" dataset was used from the Harvard Dataverse [33]. It documented the health issues that COVID-19 patients suffer from, as well as the factors that contribute to their poor prognosis. The dataset includes 582 documented cases of COVID-19-positive patients. Furthermore, age, sex, obesity, and pre-existing comorbidities have all been recognized as severe COVID-19 risk factors.

The ML classifiers applied were Bagging, J48, LR, NB, RF, SVM, and Threshold Selector to conduct three sets of experiments. Initially, we ran an experiment using all the features in the dataset to obtain a baseline accuracy and proceeded with running two further experiments with different sets of selected features based on correlation analysis. Bagging presented the highest accuracy, TPR, and FPR of 83.55%, 0.835, and 0.160, respectively, in the first experiment, which entailed utilizing nearly all dataset features. Since the models gave good results when run on both the first and second experiment, it was found that most of the features present in the dataset, namely, age, gender, BMI, infection duration, whether admitted to ICU and presence of diseases such as diabetes, cancer, hypertension, asthma, heart disease, and chronic obstructive pulmonary disease were all essential in affecting the classifier performance in detecting the mortality.

The proposed models can serve as a support system to improve decision making to detect patients at high risk of mortality. Furthermore, these models can aid in reducing the burden placed on hospitals staff by eliminating some of the routine tasks. Our study mainly explored ML models. DL models were not investigated to assess their performance. Moreover, the dataset used has a considerably small number of patients' records. Hence, for future work, we aim to acquire a larger dataset with more features including more underlying conditions to gain a greater understanding of their impact on COVID-19 patients. In addition, DL models can be applied to study their performance.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors want to thank Dr. Naya Nagy for proofreading the manuscript.

References

[1] C. I. Paules, H. D. Marston, and A. S. Fauci, "Coronavirus infections-more than just the common cold," *JAMA*, vol. 323, no. 8, pp. 707-708, 2020.

[2] N. Zhong, B. Zheng, Y. Li et al., "Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003," *The Lancet*, vol. 362, no. 9393, pp. 1353–1358, 2003.

[3] "Coronaviruses | NIH: National Institute of Allergy and Infectious diseases," 2022, https://www.niaid.nih.gov/diseases-conditions/coronaviruses.

[4] "WHO coronavirus (COVID-19) dashboard | WHO coronavirus (COVID-19) dashboard with vaccination data," 2022, https://covid19.who.int/.

[5] "Estimating mortality from covid-19," 2022, https://www.who.int/news-room/commentaries/detail/estimating-mortality-from-covid-19.

[6] Z. Wu and J. M. McGoogan, "Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China," *JAMA*, vol. 323, no. 13, pp. 1239–1242, 2020.

[7] Y. Chen, L. Ouyang, F. S. Bao et al., "A multimodality machine learning approach to differentiate severe and nonsevere COVID-19: model development and validation," *Journal of Medical Internet Research*, vol. 23, no. 4, Article ID e23948, 2021.

[8] A. Sanyaolu, C. Okorie, A. Marinkovic et al., "Comorbidity and its impact on patients with COVID-19," *SN Comprehensive Clinical Medicine*, vol. 2, no. 8, pp. 1069–1076, 2020.

[9] L. Kompaniyets, N. T. Agathis, J. M. Nelson et al., "Underlying medical conditions associated with severe COVID-19 illness among children," *JAMA Network Open*, vol. 4, no. 6, Article ID e2111182, 2021.

[10] D. García-Azorín, E. Martínez-Pías, J. Trigo et al., "Neurological comorbidity is a predictor of death in covid-19 disease: a cohort study on 576 patients," *Frontiers in Neurology*, vol. 11, 2020.

[11] F. Martos Pérez, J. Luque del Pino, N. Jiménez García et al., "Comorbilidad y factores pronósticos al ingreso en una cohorte COVID-19 de un hospital general," *Revista Clínica Española*, vol. 221, no. 9, pp. 529–535, 2021.

[12] S. Roy, S. Z. Sheikh, and T. S. Furey, "A machine learning approach identifies 5-ASA and ulcerative colitis as being linked with higher COVID-19 mortality in patients with IBD," *Scientific Reports*, vol. 11, no. 1, 2021.

[13] Y.-J. Kang, "Mortality rate of infection with COVID-19 in korea from the perspective of underlying disease," *Disaster Medicine and Public Health Preparedness*, vol. 14, no. 3, pp. 384–386, 2020.

[14] A. Banerjee, L Pasea, S Harris et al., "Estimating excess 1-year mortality associated with the COVID-19 pandemic according to underlying conditions and age: a population-based cohort study," *Lancet (London, England)*, vol. 395, no. 10238, 1715 pages, 2020.

[15] X. Guan, B. Zhang, M. Fu et al., "Clinical and inflammatory features based machine learning model for fatal risk prediction of hospitalized COVID-19 patients: results from a retrospective cohort study," *Annals of Medicine*, vol. 53, no. 1, pp. 257–266, 2021.

[16] F. Tezza, G. Lorenzoni, D. Azzolina, S. Barbar, L. A. C. Leone, and D. Gregori, "Predicting in-hospital mortality of patients with covid-19 using machine learning techniques," *Journal of Personalized Medicine*, vol. 11, no. 5, p. 343, 2021.

[17] P. Parchure, H. Joshi, K. Dharmarajan et al., "Development and validation of a machine learning-based prediction model for near-term in-hospital mortality among patients with COVID-19," *BMJ Supportive and Palliative Care*, 2020.

[18] S. Subudhi, A. Verma, A. B. Patel et al., "Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19," *NPJ Digital Medicine*, vol. 4, no. 1, 2021.

[19] 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," *Cognitive Computation*, 2021.

[20] L. Yan, H. T. Zhang, J. Goncalves et al., "An interpretable mortality prediction model for COVID-19 patients," *Nature Machine Intelligence*, vol. 2, no. 5, 2020.

[21] Y. Gao, G.-Y. Cai, W. Fang et al., "Machine learning based early warning system enables accurate mortality risk prediction for COVID-19," *Nature Communications*, vol. 11, no. 1, 2020.

[22] M. Pourhomayoun and M. Shakibi, "Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making," *Smart Health*, vol. 20, Article ID 100178, 2021.

[23] S. S. Aljameel, I. U. Khan, N. Aslam, M. Aljabri, and E. S. Alsulmi, "Machine learning-based model to predict the disease severity and outcome in COVID-19 patients," *Scientific Programming*, vol. 2021, Article ID 5587188, 10 pages, 2021.

[24] I. U. Khan, N. Aslam, M. Aljabri et al., "Computational intelligence-based model for mortality rate prediction in COVID-19 patients," *International Journal of Environmental Research and Public Health*, vol. 1812 pages, 2021.

[25] A. L. Booth, E. Abels, and P. McCaffrey, "Development of a prognostic model for mortality in COVID-19 infection using machine learning," *Modern Pathology*, vol. 34, no. 3, pp. 522–531, 2021.

[26] S. Kar, R. Chawla, S. P. Haranath et al., "Multivariable mortality risk prediction using machine learning for COVID-19 patients at admission (AICOVID)," *Scientific Reports*, vol. 11, no. 1, pp. 1–11, 2021.

[27] K. Ikemura, E. Bellin, Y. Yagi et al., "Using automated machine learning to predict the mortality of patients with COVID-19: prediction model development study," *Journal of Medical Internet Research*, vol. 23, no. 2, Article ID e23458, 2021.

[28] M. M. Banoei, R. Dinparastisaleh, A. V. Zadeh, and M. Mirsaeidi, "Machine-learning-based COVID-19 mortality prediction model and identification of patients at low and high risk of dying," *Critical Care*, vol. 25, no. 1, pp. 1–14, 2021.

[29] A. Karthikeyan, A. Garg, P. K. Vinod, and U. D. Priyakumar, "Machine learning based clinical decision support system for early COVID-19 mortality prediction," *Frontiers in Public Health*, vol. 9, Article ID 626697, 2021.

[30] 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, p. 8386, 2020.

[31] 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, no. 1, pp. 1–11, 2020.

[32] S. Li, Y. Lin, T. Zhu et al., "Development and external evaluation of predictions models for mortality of COVID-19 patients using machine learning method," *Neural Computing and Applications*, pp. 1–10, 2021.

[33] G. Santorelli, "Replication Data for: Ethnicity, Pre-existing Comorbidities, and Outcomes of Hospitalised Patients with COVID-19," 2021, https://doi.org/10.7910/DVN/RRCQEO.

[34] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, N. Rouf, and M. Mohi Ud Din, "Machine learning based approaches for detecting COVID-19 using clinical text data," *International Journal of Information Technology*, vol. 12, no. 3, pp. 731–739, 2020.

[35] A. A. T. Fernandes, D. B. Figueiredo Filho, E. C. d. Rocha, and W. d. S. Nascimento, "Read this paper if you want to learn logistic regression," *Revista de Sociologia e Política*, vol. 28, no. 74, 74 pages, 2020.

[36] R. Gandhi, *Support Vector Machine—Introduction to Machine Learning Algorithms | by Rohith Gandhi | towards Data Science*, Towards Data Science, Canada, 2018.

[37] I. Witten, E. Frank, M. Hall, and C. Pal, *Data Mining—Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Netherlands, 2016.

[38] J. Kim, J. Lee, M. Kang, and H. Sohn, "Threshold switching of Ag-Ga2Te3 selector with high endurance for applications to cross-point Arrays," *Nanoscale Research Letters*, vol. 16, no. 1, 2021.

[39] I. H. Sarker, "Machine learning: algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, no. 3, 2021.

[40] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, 2019.

[41] N. A. Mansour, A. I. Saleh, M. Badawy, and H. A. Ali, "Accurate detection of covid-19 patients based on feature correlated Naïve Bayes (FCNB) classification strategy," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 1, pp. 41–73, 2022.