Computers, Materials & Continua
DOI:10.32604/cmc.2020.012151
Article

Tech Science Press

Performance Estimation of Machine Learning Algorithms in the Factor Analysis of COVID-19 Dataset

Ashutosh Kumar Dubey[1,*], Sushil Narang[1], Abhishek Kumar[1], Satya Murthy Sasubilli[2] and Vicente García-Díaz[3]

[1]Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
[2]Workday Integration Architect Huntington, Columbus, OH, USA
[3]Department of Computer Science, University of Oviedo, Oviedo, Spain
[*]Corresponding Author: Ashutosh Kumar Dubey. Email: ashutosh.dubey@chitkara.edu.in
Received: 16 June 2020; Accepted: 25 July 2020

Abstract: Novel Coronavirus Disease (COVID-19) is a communicable disease that originated during December 2019, when China officially informed the World Health Organization (WHO) regarding the constellation of cases of the disease in the city of Wuhan. Subsequently, the disease started spreading to the rest of the world. Until this point in time, no specific vaccine or medicine is available for the prevention and cure of the disease. Several research works are being carried out in the fields of medicinal and pharmaceutical sciences aided by data analytics and machine learning in the direction of treatment and early detection of this viral disease. The present report describes the use of machine learning algorithms [Linear and Logistic Regression, Decision Tree (DT), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and SVM with Grid Search] for the prediction and classification in relation to COVID-19. The data used for experimentation was the COVID-19 dataset acquired from the Center for Systems Science and Engineering (CSSE), Johns Hopkins University (JHU). The assimilated results indicated that the risk period for the patients is 12–14 days, beyond which the probability of survival of the patient may increase. In addition, it was also indicated that the probability of death in COVID cases increases with age. The death probability was found to be higher in males as compared to females. SVM with Grid search methods demonstrated the highest accuracy of approximately 95%, followed by the decision tree algorithm with an accuracy of approximately 94%. The present study and analysis pave a way in the direction of attribute correlation, estimation of survival days, and the prediction of death probability. The findings of the present study clearly indicate that machine learning algorithms have strong capabilities of prediction and classification in relation to COVID-19 as well.

Keywords: COVID-19; linear and logistic regression; DT; KNN; SVM; SVM with grid search

1 Introduction

The World Health Organization (WHO) acknowledged novel Coronavirus Disease (COVID-19) as a pandemic on March 11, 2020, and recommended instant action predominantly for the early detection and treatment of this disease. The name, COVID-19, was suggested by WHO for the novel coronavirus that was reported to affect the lower respiratory system in the individuals in Wuhan, China [1–5]. The current name reference reported is Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) [1]. According to the report by WHO, a total of 3,247,117 COVID-19-positive cases and 229,312 COVID-19-related deaths were reported until April 30, 2020, while in India, the numbers were 33,610 cases with 1,075 deaths for the same period. The foremost symptoms of this disease are fever, tightness in the chest, unproductive cough, fatigue, breathing difficulty, lung infection, and in severe cases, pneumonia [1,2,6]. In 2020, Qiang et al. [7] reported a prediction model proposed for the detection of coronavirus infection for early alert. The data for the experimentation, which comprised protein sequences from a total of 2,666 corona virus cases, were collected from the National Genomics Data Center, China. Random forest algorithm was applied to this data, achieving an accuracy of 98.18% and Matthews Correlation Coefficient (MCC) of 0.9638. In the same year, Adhikari et al. [1] discussed the statistics of the spreading of COVID-19 throughout the world. In addition, he discussed the causes, prevention, and control mechanism for COVID-19. The major symptoms as reported for COVID-19 were fever, cough, headache, fatigue, pneumonia, diarrhea, hemoptysis, and dyspnea. Fong et al. [8] reported a critical analysis on the rapid spread of COVID-19 and its impacts. They also provided suggestions regarding the accurate forecasting mechanism. The major problems identified were limited knowledge regarding COVID-19, the uncertainty of the influencing factors, and an insufficient amount of data that is currently available. The results demonstrated that better forecasting outcomes could be achieved through the usage of polynomial neural network with corrective feedback. In 2019, Habib et al. [9] reported an analysis regarding the mortality rate and the other factors associated with the Middle East Respiratory Syndrome Coronavirus (MERS-CoV) infection. The authors discussed and reviewed the confirmed cases documented in hospital records. In the same year, Schoeman et al. [10] studied different coronaviruses along with their envelope proteins and suggested that the envelope protein is basically an integral membrane protein and is involved in the life cycle of the coronaviruses. The authors reviewed the recent progress and the current knowledge in comparison to the previous studies regarding the same. In 2020, Xu et al. [11] discussed the application of Reverse Transcription-Polymerase Chain Reaction (RT-PCR) for the early-stage detection of COVID-19. The authors also implied considering Computed Tomography (CT) imaging of the patients for the diagnosis of the disease. The segmentation was performed using a 3-dimensional (3D) deep learning model, and for experimentation, 618 CT images were selected. The overall accuracy of 86.7% was reported to be achieved. In 2020, Hassanien et al. [12] suggested a novel methodology for early-stage detection of COVID-19. The results indicated that the SVM approach followed by the authors, exhibited high accuracy in the classification of COVID-19-infected lungs. Naudé [13] discussed the role of artificial intelligence in the fight against the COVID-19 pandemic and suggested that artificial intelligence could be useful in providing early warnings, alerts, tracking, prediction, diagnosis, prognosis, treatment, and control of social contact, etc. Ma et al. [14], in their study reported in 2020, discussed the meteorological parameters in relation to COVID-19 and suggested that these were important factors as they studied the correlation between the weather parameters and the COVID-19-related deaths. The data for the meteorological parameters and air pollutants were collected from Wuhan, China, from January 20, 2020, to February 29, 2020. The authors suggested that temperature variation and humidity could affect COVID-19-related mortality. In 2020, Santosh [15] studied the role of artificial intelligence and machine learning in the decision-making process in relation to COVID-19 cases and suggested that there was a requirement of active learning-based cross-population train/test models. Hu et al. [16] discussed COVID-19 along with the causes, influencing parameters,

incidence, and prevalence of other diseases. The authors also discussed the impact of other diseases for the prognosis of COVID-19 in a patient. Roy et al. [17] conducted an online survey in order to reveal awareness, attitude, anxiety, and perceived mental health care in the context of the current pandemic situation in the Indian population. The survey received a total of 662 responses. It was inferred that there was an ardent need of spreading awareness regarding mental health issues emerging due to the pandemic. In 2020, Bullock et al. [18] discussed and reviewed the role of artificial intelligence and machine learning methodologies in relation to COVID-19. The authors also reviewed different datasets related to the same. Since machine learning appears to be an important tool in the study of the current pandemic situation, it was decided to consider it for analysis and prediction in the present study as well.

The main objectives of the present study were as follows:

- Study and analysis of the role of machine learning algorithms in relation to the COVID-19 dataset.
- Analysis and prediction of the impact of attributes and their correlation.
- Prediction of the survival status and the probability of death rates.

2 Materials and Methods

The dataset for experimentation was obtained from the data repository for the COVID-19 visual dashboard operated by the Center for Systems Science and Engineering (CSSE), Johns Hopkins University (JHU) [19]. The dataset comprised data from different countries, with attributes listed in Tab. 1. As of April 20, 2020, there were 3,397 records in the dataset. The dataset contains information regarding individual COVID-19 patients, who, after the onset of the symptoms, were admitted to a hospital and either succumbed to death or somehow recovered from this dreadful disease. In the present study, the dataset was explored to identify the presence of certain patterns among the patients. Using the dataset, the number of days the patients could survive prior to meeting their fate was determined.

Table 1: Dataset attributes discussion

S.No.	Attribute name	Discussion
1	Id	Unique case number.
2	case_in_country	An integer value for the country-wise case number.
3	Reporting date	The date on which the case was actually reported.
4	Summary	It depicts the overall case summary.
5	Location	City in which the case has been reported.
6	Country	Country in which the case has been reported.
7	Gender	Male or female.
8	Age	Age of the patient.
9	Symptom_onset	The date on which the patient started feeling the symptoms.
10	If_onset_approximated	Whether the patient was approximated after emerging of the symptoms.
11	Hosp_visit_date	The date when the patient visited the hospital.
12	International_traveler	Whether the patient has travelled internationally.
13	Domestic traveler	Whether the patient has travelled in his own country.
14	Exposure_start	The date when the patient was exposed to a COVID suspected area.

(Continued)

Table 1 (continued).

S.No.	Attribute name	Discussion
15	Exposure_end	The date when the patient left the COVID suspected area.
16	Visiting Wuhan	Whether the patient has visited Wuhan recently.
17	From Wuhan	Whether the patient belongs to Wuhan or not.
18	Death	Whether the patient has died or not.
19	Recovered	Whether the patient recovered from COVID or not.
20	Symptom and source	Description of the symptoms and source of the case information.

The following algorithms were utilized for developing a machine learning model for accurate prediction and classification from the COVID-19 dataset. Correlation analysis was performed in order to determine the dependent and independent attributes along with their strength of relationships.

2.1 Linear and Non-Linear Regression

Linear regression is nothing more than a representation of a linear model [20,21]. It demonstrates a linear relationship between x and y, where x is the input variable and y is the output variable. Therefore, it is possible to calculate y by substituting the value for x in the linear combination. The linear regression model may be presented as in Eq. (1) provided ahead:

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n \tag{1}$$

Y: The predicted value

θ_0: The intercept

$\theta_1,\ldots,\theta_n$: Model parameters

$x_1, x_2,\ldots,x_n$: Feature values(independent variables)

The simplified model may also be presented as in Eq. (2) provided ahead:

$$Y = \theta_T x_T \quad \text{where } 0 < T < n \text{ and } x_0 = 1 \tag{2}$$

When there is only one independent variable, the representation of Y prediction may be presented as the following (Eq. (3)).

$$Y = \theta_0 + \theta_1 \times x \tag{3}$$

Here, the values for θ_0 and θ_1 are selected such that the error is minimized.

In the case of only one predictor, the intercept may be calculated as presented in Eq. (4) ahead.

$$\theta_0 = \bar{y} - \theta_1 \bar{x} \tag{4}$$

$$\theta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{5}$$

This implies the following:

1. If $\theta_1 > 0$, then x and y have a positive relationship, i.e., y will increase with an increase in x.
2. If $\theta_1 < 0$, then x and y have a negative relationship, i.e., y will decrease with an increase in x.
3. If x = 0 (not included), the Eq. (5) becomes meaningless and represents beyond-the-scope value.

4. If x = 0 (included), it implies that a0 produced the average of the predicted values.
5. If θ_0 (not included), it implies that the prediction and the regression coefficient may be biased.

Our goal remains to find out optimal values of all model parameters (θ_0,…,θ_n) in order to fit the model among all data values. When there is a single independent variable or feature, the model will always shape a straight line whereas in presence of more than one feature value, the model is called a hyper plane. Many a times, the dataset shows a curvature plane instead of a straight line (when x in the Eq. (1) either has powers of more than one or in the form of any mathematical function e.g., logarithmic, exponential etc.) then a linear model may not be able to fit properly and instead a non-linear model is chosen to fit using curve fitting.

2.2 Logistic Regression

Logistic regression is used when the target variable is categorical. Therefore, logistic regression is the machine learning algorithm for classification, while linear regression is the regression algorithm for prediction.

The logistic regression model data are based on the logistic or sigmoid function:

$$g(x) = \frac{1}{1 + e^{-x}} \tag{6}$$

If it is considered in terms of Y, as in Eq. (1), as an input to the function g(x), it returns a squashed value h which ranges between 0 and 1.

$$g(y) = \frac{1}{1 + e^{-y}} \tag{7}$$

A threshold value is predicted for mapping it to a discrete class. The threshold stated below may be considered for the mapping.

$p \geq 0.5$, class = 1; $p < 0.5$, class = 0

This implies that the observation is positive if it is greater than or equal to 0.5. On the basis of the above-stated decision delimiters and logistic function, a predefined function could be generated and is presented in Eq. (8) provided ahead.

$$g(Y = 1) = \frac{1}{1 + e^{-Y}} \tag{8}$$

2.3 Decision Tree

A Decision Tree (DT) is a tree-like structure constructed for decision modeling and consequences. In a DT, the test of an attribute is denoted by the internal nodes, while the branch denotes the outcome [22,23]. The class labels are denoted by the leaf nodes. DT may be useful in various scenarios as it does not require any domain-specific knowledge for its construction. DTs are also referred to as Classification and Regression Trees (CART).

2.4 K-Nearest Neighbor

K-nearest neighbor (KNN) is a classification algorithm that is based on neighbors' majority [24]. The object assignment in a class is performed on the basis of the nearest neighbor.

Steps in the KNN algorithm:

Step 1: Data loading and initialization of K. K denotes the number of neighbors.

Step 2: Calculation and addition of the distance in ascending order.

Step 3: Selection of the first K entries from the collection sorted, as suggested in Step 2.

Step 4: Among the selected K entries, select the first K entries.

Step 5: In the case of regression, the mean of the K labels is returned, while K labels are returned in the case of classification.

2.5 Support Vector Machine

Support Vector Machine (SVM) is a classification and regression algorithm which is basically a supervised machine-learning model. SVM has been used for two-group classification problems [25]; these are hypothesis and the loss function. The main aim of SVM is to identify a hyperplane from the number of features or the N-dimension space. The purpose is to classify the data points distinctly. Therefore, for an input x, it is required to maximize the width w. This is presented in Eq. (9) provided ahead.

$$f(x) = w.x + b \tag{9}$$

$w.x + b \geq 1$, for all x of class 1

$w.x + b \leq -1$ for all x of class 2

In terms of reproducing kernel Hilbert space [26], this may be expressed as presented in Eq. (10) and (11).

$$\{f : \|f\|\_k\text{^}2 < \infty\} \tag{10}$$

$$\{f : \|f\|\_k\text{^}2 < A2\} \tag{11}$$

where ||f||_k^2 represents the reproducing kernel Hilbert space, the kernel is denoted by k, and A is a constant.

There are other hyper-parameters, for example, gamma values, which could be fine-tuned to improve the performance of the model by identifying the best combination of parameters. However, the process of identifying optimal hyper-parameter is complex. One way to approach this is to create a grid of hyper-parameters and just attempt all the possible combinations of these hyper-parameters. This method is referred to as the grid search method. This method may be useful in the over-fitting problem [27]. Both SVM and SVM with grid search were employed in the present study for classification. The complete flowchart of the proposed framework is presented in Fig. 1.

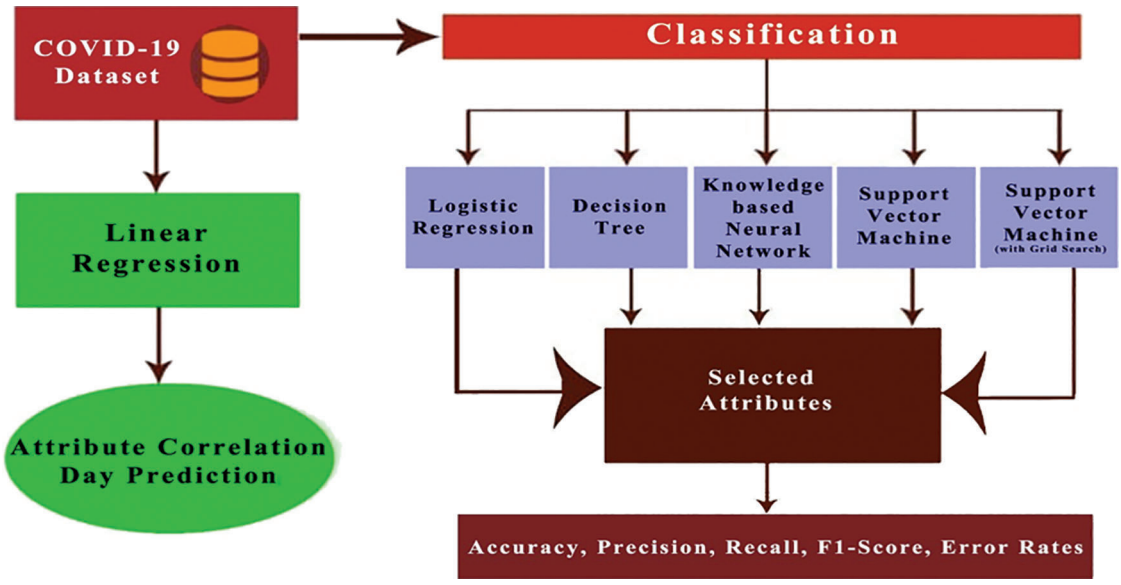

Figure 1: Flowchart of the proposed work

The following steps were performed in the above-discussed approaches.

Step 1: Data cleaning: The dataset included a few null values in certain columns, which were managed by dropping the rows corresponding to those columns, as these values could not be replaced with any other significant values.

Step 2: Feature selection: Since the dataset contained multiple columns, this step comprised the selection of those significant columns which would later be used for classification and regression analysis.

Step 3: Train–test split: The dataset was randomly divided into train data and test data. The test data considered for analysis and verification in the present study ranged between 15% and 40%.

Step 4: Fitting/Train the model: A particular classification algorithm was selected with its optimal parameters, and was used to train the model.

Step 5: Testing the model: The model was applied to the test data, and the accuracy of the model was determined by analyzing the produced confusion matrix.

Step 6: Model evaluation: The model was evaluated by calculating the error metrics.

Step 7: Model prediction: The model was applied to sample data, and the predicted value was verified.

3 Results

This section discusses the results of linear and logistic regression, DT, KNN, SVM, and SVM with Grid search applied to the above-stated dataset.

3.1 Result Based on Linear Regression

Fig. 2 depicts a grid of axes such that each of the numeric variables in the data along the y-axis is across a single row, and that along the x-axis is across a single column. The grid demonstrates the correlation among the selected independent variables. The diagonal axes present the univariate distribution of the data for the variable in that column, and fits a kernel density estimate encoding of the observations along one axis and height along the other axis. With the exception of days and age columns, all the other columns are categorical columns; therefore, the pair plot between age and days is significant in the whole pair plot. The correlation between age and death demonstrated that the maximum number of patients who died belonged to the age group of 45 and above. As the patient age increases, the number of days of survival reduces for most of the patients. It was observed that, on average, the risk period for the patients was 12–14 days, beyond which the chances of survival of the patient might increase.

Positive and negative correlation findings were deduced and are presented in Tab. 2, while Tab. 3 presents the interpretations based on these correlations. The correlation coefficient between age and death was 0.38, implying that with an increase of 1% in the age, the death probability would increase by approximately 38%. The correlation coefficient between death and Previous Illness History (PIH) was 0.79, implying that with an increase of 1% in the age, the death probability would increase by approximately 79%. Gender of the individual presented a negligible correlation with death; nonetheless, according to the data, there is a lesser probability of death in females compared to males.

On the basis of these findings, a model for bed allocation system was developed for hospitals. This model was based on x1 (prev_illness), x2 (gender_female), x3 (gender_male), and x4 (age) parameters. The regression model is presented below.

$$Y = -128x1 + 8x2 + 11x3 - 0.04x4 \tag{12}$$

The implications are presented in Fig. 3. The bed allocation system is presented in Fig. 4. The chances of survival for a female patient were high after approximately 16 days, while in the case of a male patient, the

chances of survival were high after approximately 18 days. Therefore, after this number of days specific to the gender of the patients, the patients could be discharged according to the test results. This would also be convenient in terms of bed release for admitting new patients. 'O' denotes the occupied bed and 'F' denotes the free bed. In addition, retesting from time to time is recommended. The plot between gender and days revealed that most of the male patients could survive a minimum of 8 days and a maximum of 18 days, while the minimum number of days of survival for the female patients was 3 and they could survive hardly up to 15 days. This relationship could also be explained by the other factors in the dataset, such as the presence of PIH. Another fact that could be inferred from the dataset was that the female patients constituted approximately 42% of the complete dataset, while the remaining 58% were male patients.

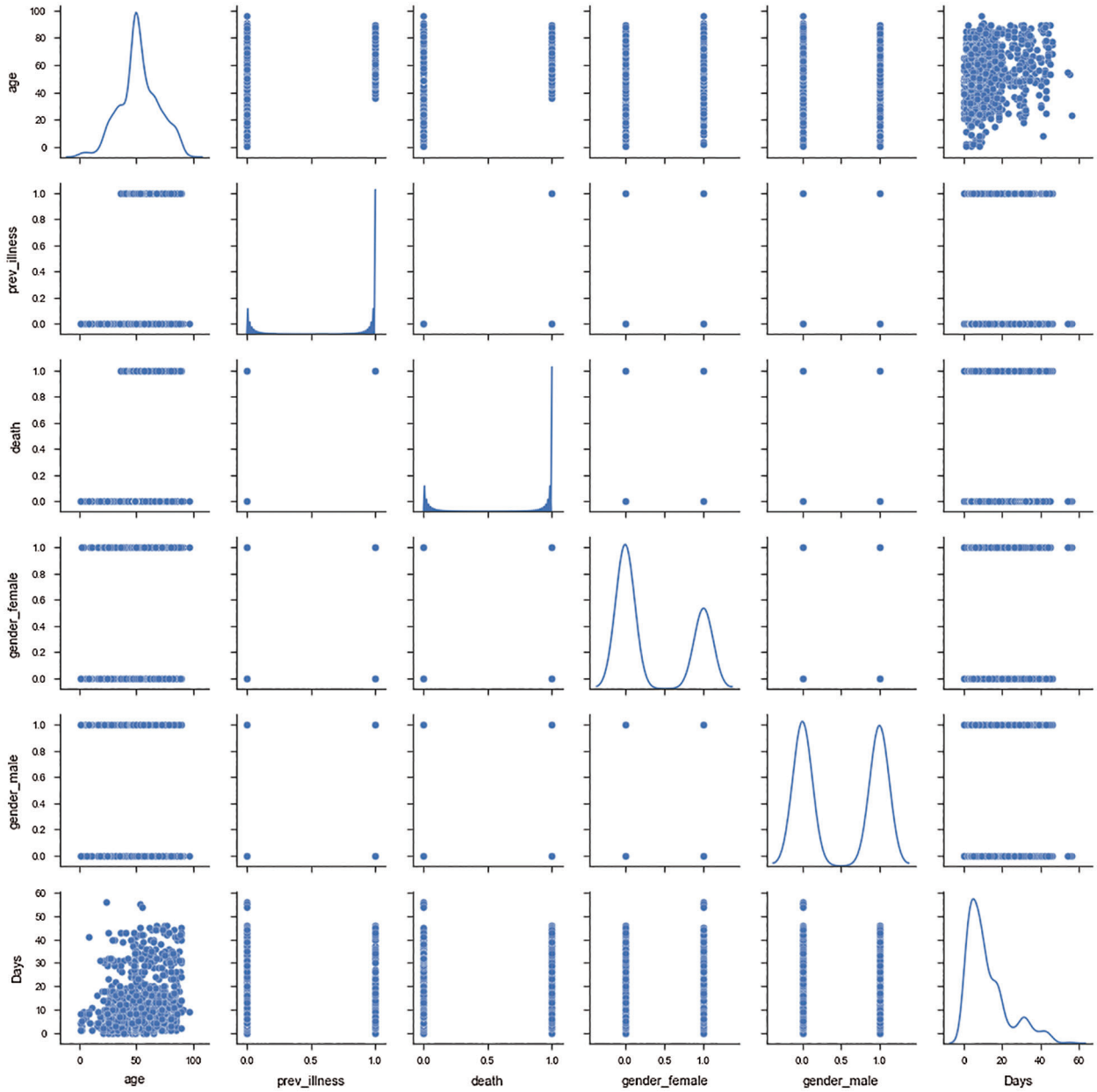

Figure 2: Correlation between different independent attributes

Table 2: Positive and negative correlation between different attributes

	Age	PIH	Death	Female	Male	Days
Age	1	0.31	0.38	0.006	0.05	0.14
PIH	0.31	1	0.79	−0.03	0.11	0.25
Death	0.38	0.79	1	−0.02	0.09	0.38
Female	0.006	−0.03	−0.02	1	−0.64	−0.006
Male	0.05	0.11	0.09	−0.64	1	−0.05
Days	0.14	0.25	0.38	−0.006	−0.05	1

Table 3: Implications of positive and negative correlation between different attributes

Positive correlation		Negative correlation	
Size of correlation	Interpretation	Size of correlation	Interpretation
0.90−1.00	Very high positive	−0.90–−1.00	Very high negative
0.70−0.90	High positive	−0.70–−0.90	High negative
0.50−0.70	Moderate positive	−0.50–−0.70	Moderate negative
0.30−0.50	Low positive	−0.30–−0.50	Low negative
0.00−0.30	Negligible correlation	−0.00–−0.30	Negligible correlation

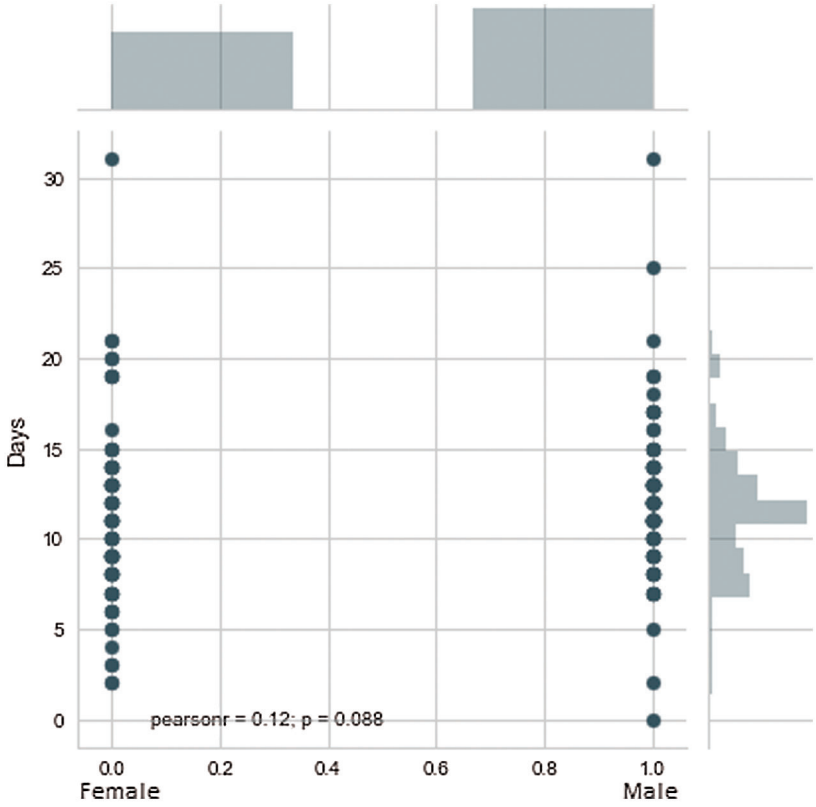

Figure 3: Model implications for the number of days in case of males and females

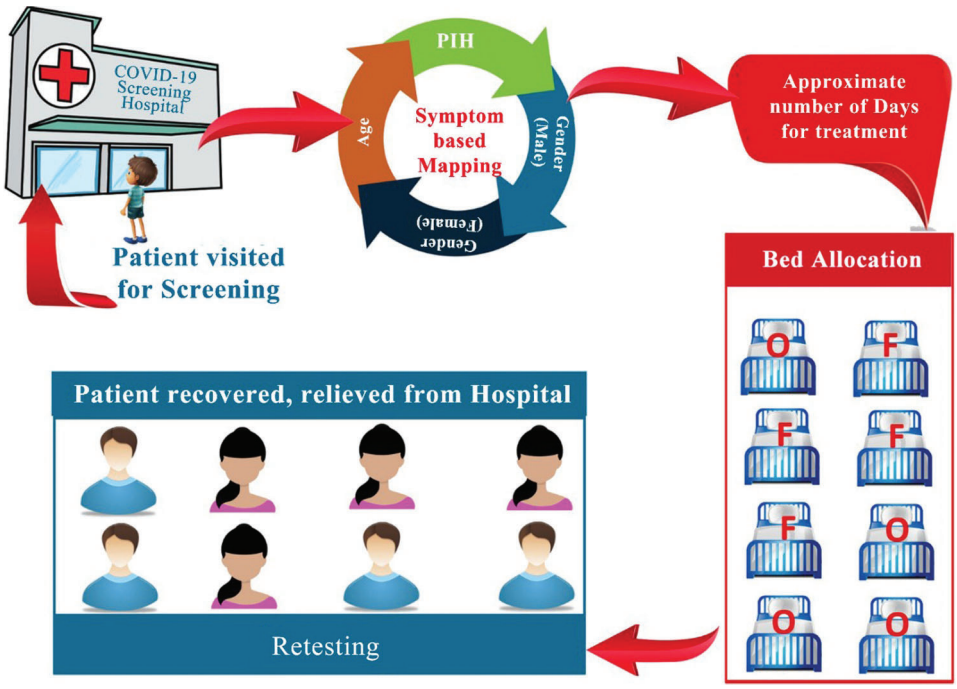

Figure 4: Bed allocation system based on the regression model

Fig. 5 presents the predicted value for days when the sample inputs were fed to the proposed model. The sample inputs comprised the patients' PIH, their gender (Male or Female), and their age. It was indicated that the patients of COVID-19 exhibited the highest number of days for the treatment when there was no PIH. In addition, it was observed that the approximate number of days for the treatment was reduced for older individuals. Both these findings, when taken together, imply that the older patients who also have PIH would exhibit lesser number of days for the treatment. This also implies that the chances of survival for such patients were less. Another inference from these results comes out that female patients exhibit better chances of survival compared to male patients.

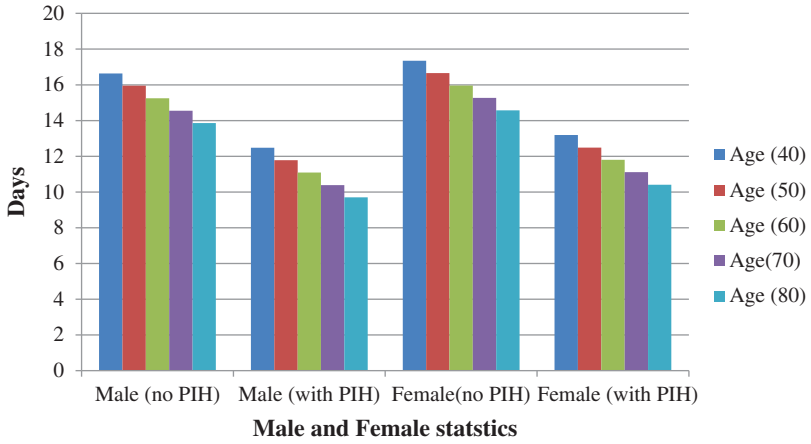

Figure 5: Approximate number of days of treatment in case of males and females (with and without PIH)

3.2 Model Accuracy and Evaluation

The performance of the model was evaluated on the basis of the following measures:

Accuracy: The ratio of predicted implications to total implications.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{13}$$

where, TP denotes true positive, TN represents true negative, FP represents false positive, and FN denotes false negative.

Precision: The ratio of predicted positive implications to total positive implications.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{14}$$

Recall (Sensitivity): The ratio of predicted positive implications to all implications in the actual class—yes.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{15}$$

F1-Score: This score is calculated on the basis of precision and the recall weighted average.

$$\text{F1 Score} = \frac{2 \times (recall \times precision)}{recall + precision} \tag{16}$$

Here, 0 and 1 represent the live status and dead status. Tabs. 4–8 present the values for accuracy, precision, recall, and F-score obtained using logistic regression, DT, KNN, SVM, and SVM with Grid search. The test data considered here were 15%–40%. In the case of KNN, different values of K were considered running for n times for the selection of K, for which the error was minimized. The mean values for the count of false predictions in each prediction performed by the classifier were recorded as error rates and were visualized (Fig. 6). The value of K was also cross-verified using K-means cross-validation. As evident from the figure (Fig. 6), the error rate was lowest for k = 15, which was eventually selected for the classifier. Fig. 7 presents the comparison of all the five models in terms of predicting the chances of death for a COVID-19 patient on the basis of the input characteristics of the patient. The same random seed was selected for all the algorithms for a fair comparison. A comparison of the average accuracy values among all the models was performed and plotted accordingly. It is evident that SVM with Grid search presented the highest accuracy of approximately 95%, followed by the decision tree with an accuracy of approximately 94%.

Table 4: Accuracy obtained based on logistic regression

Test data (%)	Accuracy	Precision		Recall		F-score	
		0	1	0	1	0	1
40	0.94	0.93	0.99	1.00	0.73	0.96	0.84
35	0.93	0.92	0.99	1.00	0.70	0.96	0.82
30	0.94	0.93	0.98	1.00	0.73	0.96	0.84
25	0.93	0.92	0.98	1.00	0.70	0.96	0.81
20	0.93	0.92	0.97	1.00	0.67	0.95	0.80
15	0.94	0.93	0.97	0.99	0.74	0.96	0.84

Table 5: Accuracy obtained based on DT

Test data (%)	Accuracy	Precision		Recall		F-score	
		0	1	0	1	0	1
40	0.94	0.94	0.97	0.99	0.77	0.96	0.86
35	0.94	0.93	0.96	0.99	0.75	0.96	0.84
30	0.95	0.95	0.96	0.99	0.80	0.97	0.87
25	0.94	0.94	0.96	0.99	0.77	0.97	0.85
20	0.95	0.95	0.94	0.99	0.80	0.97	0.86
15	0.94	0.94	0.98	0.98	0.79	0.96	0.85

Table 6: Accuracy obtained based on KNN

Test data (%)	Accuracy	Precision		Recall		F-score	
		0	1	0	1	0	1
40	0.88	0.87	0.96	1.00	0.45	0.93	0.61
35	0.88	0.87	0.92	0.99	0.48	0.93	0.63
30	0.89	0.89	0.90	0.98	0.55	0.93	0.69
25	0.90	0.90	0.91	0.98	0.59	0.94	0.72
20	0.91	0.90	0.92	0.99	0.62	0.94	0.74
15	0.90	0.90	0.90	0.98	0.62	0.94	0.73

Table 7: Accuracy obtained based on SVM

Test data (%)	Accuracy	Precision		Recall		F-score	
		0	1	0	1	0	1
40	0.88	0.87	0.96	1.00	0.45	0.93	0.61
35	0.88	0.87	0.92	0.99	0.48	0.93	0.63
30	0.89	0.89	0.90	0.98	0.55	0.93	0.69
25	0.90	0.90	0.91	0.98	0.59	0.94	0.72
20	0.91	0.90	0.92	0.99	0.62	0.94	0.74
15	0.90	0.90	0.90	0.98	0.62	0.94	0.73

Root Mean-Squared Error (RMSE) and Mean Absolute Scaled Error (MASE) were used for model error rate evaluation. RMSE demonstrates underestimation and overestimation within the same pattern, while MASE is beneficial for relative accuracy. A lower value of error implies a better fit.

Table 8: Accuracy obtained based on SVM with Grid search

Test data (%)	Accuracy	Precision		Recall		F-score	
		0	1	0	1	0	1
40	0.94	0.93	0.99	1.00	0.74	0.96	0.85
35	0.94	0.93	0.99	1.00	0.73	0.96	0.84
30	0.95	0.94	0.98	1.00	0.76	0.97	0.86
25	0.94	0.93	0.98	1	0.72	0.96	0.83
20	0.93	0.93	0.97	1.00	0.71	0.96	0.82
15	0.95	0.94	0.97	0.99	0.79	0.97	0.87

Figure 6: Error rate and K value in case of KNN

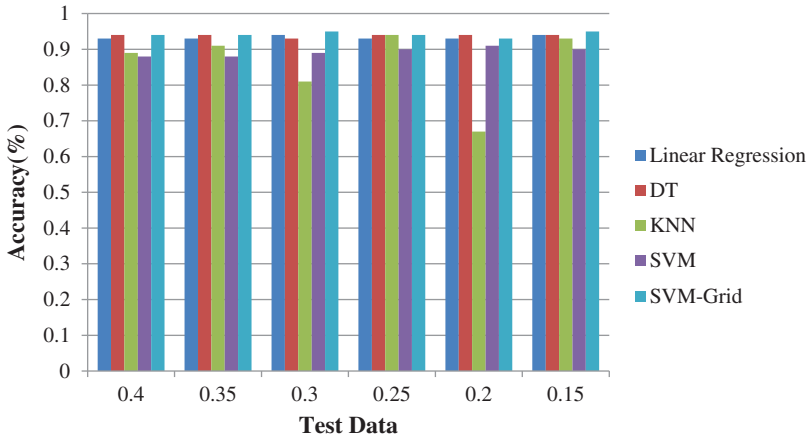

Figure 7: Comparison of algorithms

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\widehat{y_i} - y_i)^2}{n}} \tag{17}$$

$\widehat{y_i}$ = Predicted values

y_i = Observed values

n = Number of observations

$$\text{MASE} = \frac{1}{n} \sum_{i=1}^{n} \frac{|f_i - a_i|}{\sum_{j=2}^{n} |a_i - a_{j-1}|} \tag{18}$$

a_i = Actual time series

f_i = Forecast results

The error rate presented in Fig. 8 also suggests that SVM with Grid search was the most efficient, followed by DT.

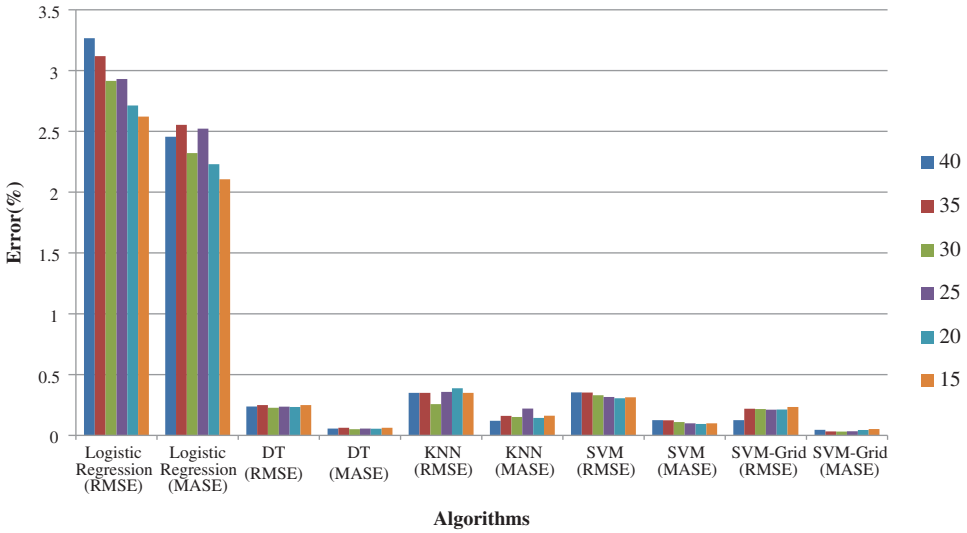

Figure 8: Logistic regression, DT, KNN, SVM and SVM-Grid based error rate comparison

4 Discussion

The key findings of the present study are as follows:

1. The results indicated that, on average, the risk period for the patients was 12–14 days, beyond which the chances of survival of the patient may increase.
2. The main attributes affecting the prediction and classification were PIH, gender, and the number of days.
3. The chances of survival for a female patient were high after approximately 16 days, while in the case of a male patient, the chances of survival were high after approximately 18 days.
4. The results indicated that the patients of COVID-19 exhibited the highest number of days for the treatment when there was no PIH. In addition, female patients were indicated to have a higher number of days for treatment compared to male patients.

5. It was inferred that even though the virus could attack males mostly, probably because males were more exposed to the virus because of being at places of work or outside of their homes to earn their livelihood, nonetheless, the immunity level of male patients was higher than that of the female patients.
6. Finally, it was indicated that machine learning algorithms were capable of providing prediction and classification in relation to COVID-19 as well. SVM with Grid search has observed to be the most efficient in this regard, followed by DT.

4.1 Limitations

Despite promising results, the present study also has certain limitations. The first one remains the limited availability of data related to COVID-19 patients. The information regarding the attributes varies continually, and updation in the dataset is quite possible in the future. Therefore, few results may vary accordingly. However, that would not critically alter the impact of the important attributes, such as age, PIH, and gender, as these have already been considered individually as well as in association with other attributes. Also, if the number of countries, affected by the COVID-19 virus, increase, results may vary further. Currently, the impact is limited, as the country's association is limited. Finally, it must be stated that further research is required to elucidate how a hybrid combination of different computational methods could be formed and used in order to provide the most effective outcomes when used in combination.

5 Conclusion

The present report discusses the current scenario due to the COVID-19 pandemic in terms of the current statistics, the major symptoms exhibited by the patients, and its impact on the survival and the death rates. The machine learning algorithms, such as linear as well as non-linear regression, logistic regression, DT, KNN, SVM, and SVM with Grid search, have been considered for the purposes of classification and prediction in relation to COVID-19. The present report discusses three main attributes: PIH, gender, and the number of days. The dataset considered in the present study was obtained from CSSE, JHU. The results of experimentation indicated that the chances of death are higher in the case of PIH and patients older in age. On the basis of these attributes, the average number of treatment days after which the chances of survival may increase for the patient could be calculated. In addition, it was indicated that females exhibit lesser chances of death compared to males. The SVM with Grid search outperformed all the other algorithms studied, in terms of classification accuracy.

Funding Statement: The author(s) received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. P. Adhikari, S. Meng, Y. Wu, Y. Mao, R. Ye *et al.,* "Epidemiology, causes, clinical manifestation and diagnosis, prevention and control of coronavirus disease (COVID-19) during the early outbreak period: A scoping review," *Infectious Diseases of Poverty*, vol. 9, no. 1, pp. 1–12, 2020.

[2] X. Jiang, M. Coffee, A. Bari, J. Wang, X. Jiang *et al.,* "Towards an artificial intelligence framework for data-driven prediction of Coronavirus clinical severity," *Computers, Materials & Continua*, vol. 63, no. 1, pp. 537–551, 2020.

[3] Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou *et al.,* "Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia," *New England Journal of Medicine*, vol. 382, no. 13, pp. 1199–1207, 2020.

[4] CDC, "Novel coronavirus, Wuhan, China," 2020. [Online]. Available: https://www.cdc.gov/coronavirus/2019-nCoV/summary.html.

[5] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang *et al.,* "A novel coronavirus from patients with pneumonia in China, 2019," *New England Journal of Medicine*, vol. 382, no. 8, pp. 727–733, 2020.

[6] F. Wu, S. Zhao, B. Yu, Y. M. Chen, W. Wang *et al.,* "A new coronavirus associated with human respiratory disease in China," *Nature*, vol. 579, no. 7798, pp. 265–269, 2020.

[7] X. L. Qiang, P. Xu, G. Fang, W. B. Liu and Z. Kou, "Using the spike protein feature to predict infection risk and monitor the evolutionary dynamic of coronavirus," *Infectious Diseases of Poverty*, vol. 9, no. 1, pp. 1–8, 2020.

[8] S. J. Fong, G. Li, N. Dey, R. G. Crespo and E. Herrera-Viedma, "Finding an accurate early forecasting model from small dataset: a case of 2019-ncov novel coronavirus outbreak," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 1, pp. 132–140, 2020.

[9] A. M. Habib, M. A. Ali, B. R. Zouaoui, M. A. Taha and B. S. Mohammed, "Clinical outcomes among hospital patients with Middle East respiratory syndrome coronavirus (MERS-CoV) infection," *BMC Infectious Diseases*, vol. 19, no. 1, pp. 1–6, 2019.

[10] D. Schoeman and B. C. Fielding, "Coronavirus envelope protein: Current knowledge," *Virology Journal*, vol. 16, no. 1, pp. 1–22, 2019.

[11] X. Xu, X. Jiang, C. Ma, P. Du, X. Li *et al.,* "A deep learning system to screen coronavirus disease 2019 pneumonia," *Engineering*, pp. 1–8, 2020.

[12] A. E. Hassanien, L. N. Mahdy, K. A. Ezzat, H. H. Elmousalami and H. A. Ella, "Automatic x-ray COVID-19 lung image classification system based on multi-level thresholding and support vector machine," *MedRxiv*, [Preprint], 2020.

[13] W. Naudé, "Artificial intelligence against COVID-19: An early review," 2020. [Online]. Available: https://www.econstor.eu/bitstream/10419/216422/1/dp13110.pdf.

[14] Y. Ma, Y. Zhao, J. Liu, X. He, B. Wang *et al.,* "Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China," *Science of the Total Environment*, vol. 724, no. 1, pp. 1–7, 2020.

[15] K. C. Santosh, "AI-driven tools for coronavirus outbreak: Need of active learning and cross-population train/test models on multitudinal/multimodal data," *Journal of Medical Systems*, vol. 44, no. 5, pp. 1–5, 2020.

[16] Y. Hu, J. Sun, Z. Dai, H. Deng, X. Li *et al.,* "Prevalence and severity of corona virus disease 2019 (COVID-19): A systematic review and meta-analysis," *Journal of Clinical Virology*, vol. 127, pp. 1–7, 2020.

[17] D. Roy, S. Tripathy, S. K. Kar, N. Sharma, S. K. Verma *et al.,* "Study of knowledge, attitude, anxiety & perceived mental healthcare need in Indian population during COVID-19 pandemic," *Asian Journal of Psychiatry*, vol. 51, pp. 1–7, 2020.

[18] J. Bullock, K. H. Pham, C. S. Lam and M. Luengo-Oroz, "Mapping the landscape of artificial intelligence applications against COVID-19," *arXiv preprint, arXiv:2003.11336*, 2020.

[19] https://www.kaggle.com/anjanatiha/corona-virus-time-series-dataset.

[20] S. Sperandei, "Understanding logistic regression analysis," *Biochemia Medica*, vol. 24, no. 1, pp. 12–18, 2014.

[21] J. C. Stoltzfus, "Logistic regression: A brief primer," *Academic Emergency Medicine*, vol. 18, no. 10, pp. 1099–1104, 2011.

[22] Y. Zhong, "The analysis of cases based on decision tree," in *Proc. IEEE, ICSESS*, Beijing, China, pp. 142–147, 2016.

[23] H. Yang and S. Fong, "Optimized very fast decision tree with balanced classification accuracy and compact tree size," in *Proc. IEEE, ICMIA*, Macao, China, pp. 57–64, 2011.

[24] Z. Chen, L. J. Zhou, L. X. Da, J. N. Zhang and W. J. Huo, "The lao text classification method based on KNN," *Procedia Computer Science*, vol. 166, pp. 523–528, 2020.

[25] G. N. Kouziokas, "SVM kernel based on particle swarm optimized vector and Bayesian optimized SVM in atmospheric particulate matter forecasting," *Applied Soft Computing*, vol. 93, pp. 1–18, 2020.

[26] M. J. Safari, S. R. Arashloo and A. D. Mehr, "Rainfall-runoff modeling through regression in the reproducing kernel Hilbert space algorithm," *Journal of Hydrology*, vol. 587, pp. 1–12, 2020.

[27] S. Mutasa, S. Sun and R. Ha, "Understanding artificial intelligence based radiology studies: What is overfitting," *Clinical Imaging*, vol. 65, pp. 96–99, 2020.