20

Forecast and prediction of COVID-19 using machine learning

Deepak Painuli[1], Divya Mishra[2], Suyash Bhardwaj[1], Mayank Aggarwal[1]

[1]GURUKUL KANGRI VISHWAVIDYALAYA, HARIDWAR, UTTARAKHAND, INDIA;
[2]UTTARAKHAND TECHNICAL UNIVERSITY, DEHRADUN, UTTARAKHAND, INDIA

1. Introduction

In this era of automation, artificial intelligence and data science have important role in the health care industry. These technologies are so well-connected that medical professionals can easily manage their roles and patient care. All health care organizations work hard to develop an automated system that can be used to accept the challenges faced in health care. Scientists are working on machine learning (ML) to develop smart solutions to diagnose and treat disease. ML is capable of detecting disease and virus infections more accurately so that patients' disease can be diagnosed at an early stage, the dangerous stages of diseases can be avoided, and there can be fewer patients. In the same manner, ML can be used to automate the task of predicting COVID-19 infection and help forecast future infection tallies of COVID-19. In this chapter, we include methods for forecasting future cases based on existing data. ML approaches are used and two solutions, one for predicting the chances of being infected and other for forecasting the number of positive cases, are discussed. A trial was done for different algorithms, and the algorithm that gave the results with the best accuracy is covered in the chapter. The chapter discusses autoregressive integrated moving average (ARIMA) time series for forecasting confirmed cases for various states in India. Two classifiers, random forest and extra tree classifier (ETC), are selected; both have an accuracy of more than 90%. Of the two, ETC has 93.62% accuracy. These results can be used to take corrective measures by different government bodies. The availability of techniques for forecasting infectious disease can make it easier to fight against infectious disease such as COVID-19.

The objective of the chapter is to find the best-performing ML model for predicting and forecasting COVID-19. Afterward, reader will obtain a glimpse of some ML fundamentals and how ML can be used to predict and forecast COVID-19, which may help in future health care automation tasks using ML and data science.

Data Science for COVID-19. https://doi.org/10.1016/B978-0-12-824536-1.00027-7

The chapter is divided into eight sections. Section 2 introduces COVID-19, the incubation period of COVID-19, and other details about COVID-19. Section 3 gives a brief overview of ML and its methods. Section 4 describes how ML can be used in COVID-19. Section 5 describes different ML techniques for prediction and forecasting, including a general ML process flowchart. Sections 6 and 7 describe the proposed symptoms-based prediction model for classification of COVID-19 infection and the ARIMA model for forecasting the future confirmed case count of COVID-19 in India. Section 8 focuses on conclusions and future work.

2. Introduction to COVID-19

COVID-19 is not just a name now. It has become a deadly widespread virus that has affected tens of thousands of people all over the world. Its origin was Wuhan City, China in Dec. 2019. When people were unaware of the virus, COVID-19 started to spread from one person to another; it has slowly reached almost all countries and has become a pandemic [1–3].

COVID-19 is the short form for coronavirus disease 2019, an illness caused by a novel coronavirus (nCoV) now known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2); formerly called 2019-nCoV. COVID-19 was not the formal name of this virus; it was called SARS-CoV-2 by the International Committee on Taxonomy of Viruses because its symptoms were related to the virus that caused the SARS outbreak in 2003. However, this virus had not previously appeared in humans, and this time, they were severely infected by the virus, so to avoid confusion with other viruses, the World Health Organization (WHO) named it COVID-19 to communicate with the public [2–5].

During its early stages, COVID-19 was first identified as only an outbreak of respiratory illness cases in Wuhan City, Hubei Province, China. On Dec. 31, 2019, China reported about this respiratory disease to the WHO. It was declared to be COVID-19, a global health emergency, by the WHO on Jan. 30, 2020. According to records of WHO, in 2009, H1N1 was declared to be a global pandemic after which, on Mar. 11, 2020, COVID-19 was declared a global pandemic by the WHO [2].

The name COVID-19 was selected because the WHO does not want to associate the origins of the virus in terms of populations, geography, or animals to cause stigma [5].

According to the WHO and other health agencies, coronaviruses are defined as a collection of viruses whose symptoms ranges from the common cold to more severe diseases. However, nCoV is a new type of virus not been previously seen in humans.

Countries across the globe quickly identified this respiratory disease as the cases of COVID-19 rapidly increased. More and more people were infected with COVID-19 since the day it was identified in China [1,3]. Since it was declared as the pandemic, the WHO has published guidance regarding this virus for all countries, including how the people may identify whether they are infected by this disease, how to remain unaffected by the virus, what kind of precautions should be taken care, when to go to the hospital, levels of conditions of people who are infected, and symptoms of this virus after a deep examination of infected people [2–5].

The WHO continuously shares information with people in different countries about this virus so that the public does not panic. During the early days of COVID-19, the WHO did not suggest avoiding travel. Strict suggestions were to distance from infected persons, wash hands regularly, and, if experiencing coughing or a cold, covering the mouth.

However, later on, travel history became one of the important identifiers of COVID-19, and based on this information, screening of all persons traveling from different countries, especially from infected areas, was done regularly. All persons coming from other countries were recommended to be isolated at home for around 14 days, because that was the symptomatic period of this virus, as mentioned by the WHO. If a person showed any symptoms of illness, he or she was taken to the hospital for treatment [2].

2.1 Incubation period of COVID-19

The incubation period is the time between when someone catches the virus and when symptoms start to appear. As reported by the WHO, this virus has an incubation period of 2—14 days in the human body [4,6].

According to the Centers for Disease Control and Prevention (CDC), mild symptoms of the virus start appearing within 5 days and become worse afterward [7].

However, more recent data on patients showed that the incubation period had increased from 14 to 20 or 28 days as the virus started mutating, and after many negative tests, it suddenly revealed a positive result.

It was reported that patients have tested positive for the virus without having symptoms owing to a strong immunity system. As, the symptoms of this virus do not appear with strong immunity system, so if we come in contact with the person affected by the virus but with strong immunity system then we can definitely get infected [4,8,9].

2.2 How it is transmitted

The coronavirus is transmitted from person to person when they are directly in contacted with each other or when the infected person sneezes or coughs. It is a respiratory disease, so it directly affects the respiratory system [2]. According to CDC, nCoV is reported to be highly contagious, which means it spreads easily from among persons. [6]. It can also be spread when a person touches a surface or edible items that have come into contact with an infected person.

2.3 Symptoms of COVID-19

The most common symptoms of COVID-19 are coughing and sneezing, fever, and breathing problems. In addition to these symptoms, diarrhea, hearing problems, a loss of sense of smell, chest pains, and nasal congestion are experienced.

The WHO released precautionary measures to avoid infection from COVID-19 virus. They include covering the face with a mask or cloth, avoid handshaking and instead bowing with namaste, following social distancing, and enforcing a lockdown [1–5].

2.4 Countries most affected by COVID-19

Several countries are affected by COVID-19. In some of those most affected, many thousands of people have died from the coronavirus [10–19]:

(1) China reported its first positive case appeared in Wuhan, China, which is also the origin of COVID-19, on Dec. 31, 2019. An update of COVID-19 in China cited 82,816 cases out of which 4632 died and 77,346 recovered and 838 cases of which remained active. [11,12]; (2) Italy reported its first positive case on Jan. 31, 2020 in Rome. There, of 192,994 cases, 25,969 died, 60,498 recovered, and 106,527 remained active [13]; (3) Spain reported its first positive case on Jan. 31, 2020. There were 223,759 cases, out of which 22,902 died, 95,708 recovered, and 105,149 remained active [14]; (4) the United States reported its first case on Jan. 19, 2020. There were 928,364 cases, out of which 52,356 died and 110,490 recovered. It is a large number. [15]; (5) Germany reported its first case on Jan. 27, 2020. There were 155,407 cases, out of which 5802 died, 109,800 recovered, and 39,805 remained active [16]; (6) France reported its first case on Jan. 24, 2020. There were 159,828 cases, out of which 22,245 died, 43,493 recovered, and 94,090 remained active. [17]; (7) Iran reported its first case on Feb. 19, 2020. There were 89,328 cases, out of which 5650 died, 68,193 recovered, and 15,485 cases remained active [18]; (8) the United Kingdom reported its first case on Jan. 29, 2020. There were 148,377 cases, out of which 20,319 died [15]; (9) Turkey reported its first case on Mar. 10, 2020. There were 104,912 cases, out of which 2600 died and 21,737 recovered [19].

Besides all of these countries, the first case in India was reported on Jan. 30, 2020 in Kerala. There were 24,942 cases, out of which 780 died and 5498 recovered.

3. Introduction to machine learning

According to Arthur Samuel (1959), ML is the field of study that gives computers the ability to learn without being explicitly programmed. Thus, we can define ML as the field of computer science in which machines can be designed that can program themselves [20].

The process of learning is simply learning from experience or observations from previous work, such as examples, or instruction, to look for patterns in data and with the help of examples, provided the system can make better decisions. The basic aim of ML is to make computers learn automatically with no human intervention and to adjust perform actions accordingly [20,21].

Fig. 20.1 shows the process of ML. Past data are used to train the model, and then this trained model is used to test new data and then for prediction. The trained ML model's performance is evaluated using some portion of available past data (which is not present

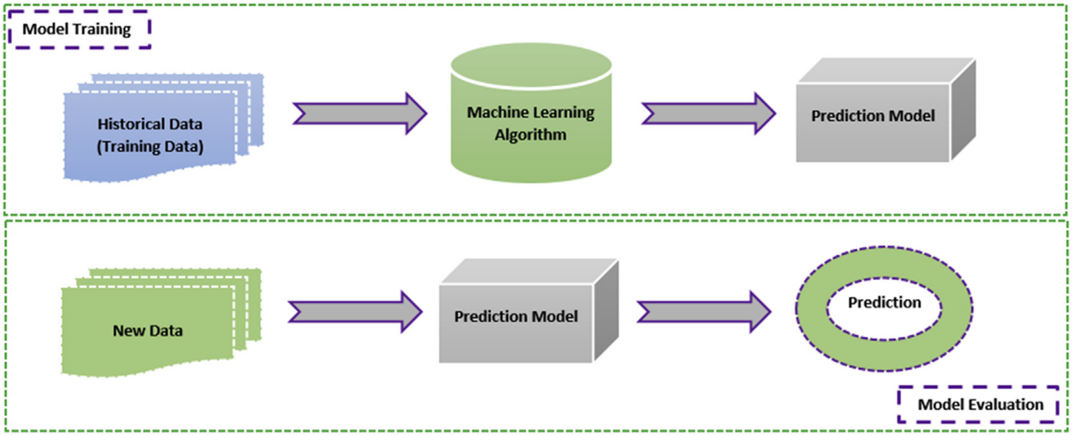


FIGURE 20.1 Machine learning process.

during training). This is usually referred as the validation process. In this process, the ML model is evaluated for its performance measure, such as accuracy. Accuracy describes the ML model's performance over unseen data in terms of the ratio of the number of correctly predicted features and total available features to be predicted.

3.1 Some machine learning methods

ML algorithms can be divided into supervised or unsupervised learning:

(1) Supervised ML algorithms is a type of ML technique that can be applied according to what was previously learned to get new data using labeled data and to predict future events or labels. In this type of learning, supervisor (labels) is present to guide or correct. For this first analysis, the known training set and then the output values are predicted using the learning algorithm. The output defined by the learning system can be compared with the actual output; if errors are identified, they can be rectified and the model can be modified accordingly [20].

(2) Unsupervised ML algorithms: In this type, there is no supervisor to guide or correct. This type of learning algorithm is used when unlabeled or unclassified information is present to train the system. The system does not define the correct output, but it explores the data in such a way that it can draw inferences (rules) from datasets and can describe hidden structures from unlabeled data [20–22].

(3) Semisupervised ML algorithms are algorithms that are between the category of supervised and unsupervised learning. Thus, this type of learning algorithm uses both unlabeled and labeled data for training purposes, generally a small amount of labeled data and a large amount of unlabeled data. This type of method is used to improve the accuracy of learning [20–22].

(4) Reinforcement ML algorithms is a type of learning method that gives rewards or punishment on the basis of the work performed by the system. If we train the system to perform a certain task and it fails to do that, the system might be punished; if it performs perfectly, it will be rewarded. It typically works on 0 and 1, in which 0 indicates a punishment and 1 indicates a reward.

It works on the principle in which, if we train a bird or a dog to do some task and it does exactly as we want, we give it a treat or the food it likes, or we might praise it. This is a reward. If it did not perform the task properly, it might be scolded as a punishment by us. [20–22].

4. Use of machine learning in COVID-19

ML is used in various fields, including medicine to predict disease and forecast its outcome. In medicine, the right diagnosis and the right time are the keys to successful treatment. If the treatment has a high error rate, it may cause several deaths. Therefore, researchers have started using artificial intelligence applications for medical treatment. The task is complicated because the researchers have to choose the right tool: it is a matter of life or death [23].

For this task, ML achieved a milestone in the field of health care. ML techniques are used to interpret and analyze large datasets and predict their output. These ML tools were used to identify the symptoms of disease and classify samples into treatment groups. ML helps hospitals to maintain administrative processes and treat infectious disease [24–26].

ML techniques were previously used to treat cancer, pneumonia, diabetes, Parkinson disease, arthritis, neuromuscular disorders, and many more diseases; they give more than 90% accurate results in prediction and forecasting [22,23].

The pandemic disease known as COVID-19 is a deadly virus that has cost the lives of many people all over the world. There is no treatment for this virus. ML techniques have been used to predict whether patients are infected by the virus based on symptoms defined by WHO and CDC [2,6].

ML is also used to diagnose the disease based on x-ray images. For instance, chest images of patients can be used to detect whether a patient is infected with COVID-19 [25,26].

Moreover, social distancing can be monitored by ML; with the help of this approach, we can keep ourselves safe from COVID-19 [2,3,24].

5. Different techniques for prediction and forecasting

Various ML techniques are used to predict and forecast future events. Some ML techniques used for prediction are support vector machine, linear regression, logistic regression, naive Bayes, decision trees (random forest and ETC), K-nearest neighbor, and neural networks (multilayer perceptron) [27–29].

Similarly, some ML techniques used to forecast future events are naive approach, moving average, simple exponential smoothing, Holt's linear trend model, Holt-Winters model, Seasonal Autoregressive Integrated Moving Average Exxogenous Model (SARIMAX) and Autoregressive Integrated Moving Average Model (ARIMA).

Each technique has unique features and is used differently based on the accuracy results. The model with the best accuracy during the model evaluation process is chosen for prediction or forecasting. In the same way, we identified and used the ETC for the symptom-based prediction of COVID-19 and the ARIMA forecasting model to forecast the number of confirmed cases of COVID-19 in India, because they had the best accuracy results among all classifier and forecasting methods we used when we evaluated model performance.

Fig. 20.2 shows a flowchart of the ML process. It defines how data are collected and preprocessed, and then are divided into a training dataset and test dataset for training and performance evaluation.

6. Proposed method for prediction

A symptom-based predictive model was proposed to predict COVID-19 based on symptoms defined by the WHO and CDC [1–3,6].

Because there is no proper description of symptoms declared by the WHO, based on some existing symptoms, we defined a model used to predict the disease according to the accuracy given by the model [1,2].

We created a symptom database in which rules were created and used as input. Then, these data were used as raw data. Then, feature selection took place as part of preprocessing data. The data were divided into training data (80% of data) and test data (20% of data), usually known as the train-test split process. This split is generally done in a stratified or random manner so that population distribution in both groups consists of shuffled data, which leads minimized bias or skewness in the data. Training data were used to train the ML classifier that we used in the model, and test data were used to test that classifier in terms of accuracy received over a predefined unseen portion of the dataset [29–39].

In our work, the symptoms and patient's class dataset was defined on the basis of symptoms such as fever, cough, and sneezing, whether the patient had traveled to an infected place, age, and whether the patient had a history of disease that could increase the possibly of being infected by the virus [1,2].

This dataset was then further divided into two sets (training set and testing set) using the test-train split method. The system was trained on the basis of training set data and the accuracy of the ML classifier, and then evaluated over the testing set. Finally, the model was used to predict the probability of infection from the disease using new patient data in terms of positive or negative [34,35].

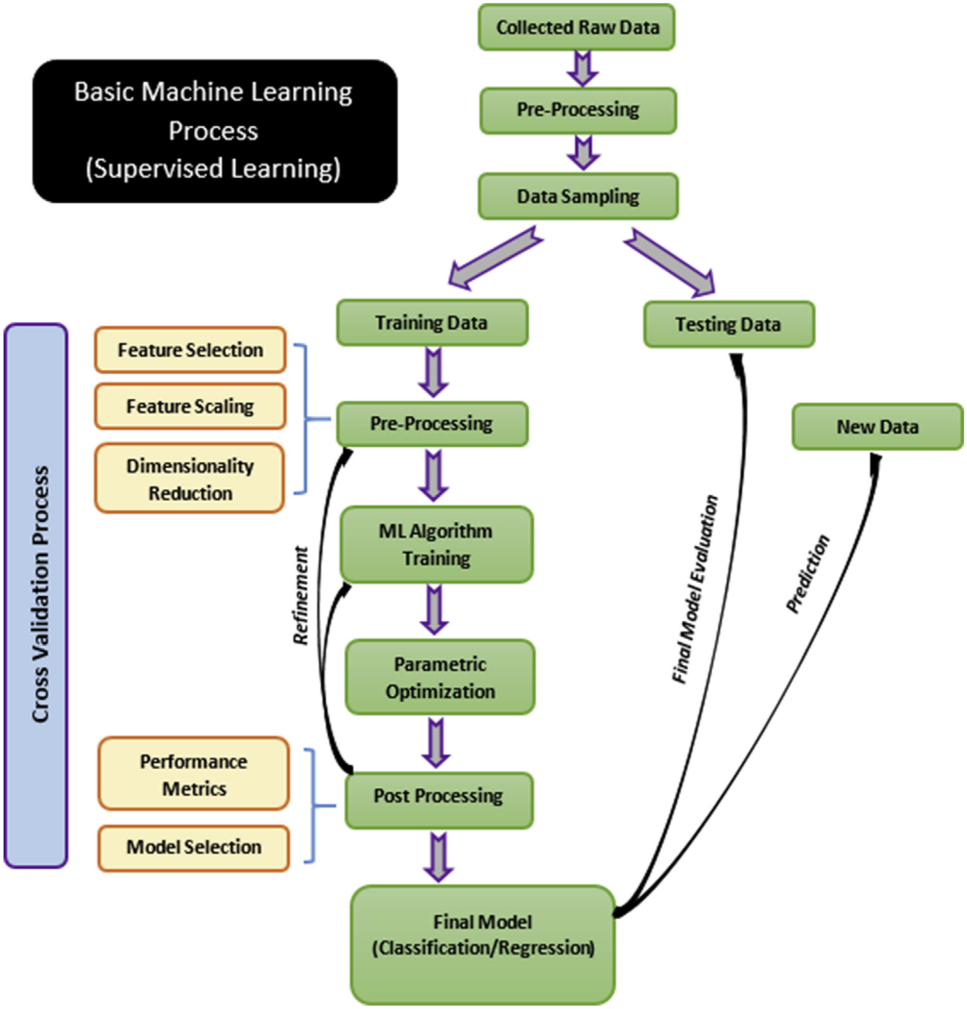
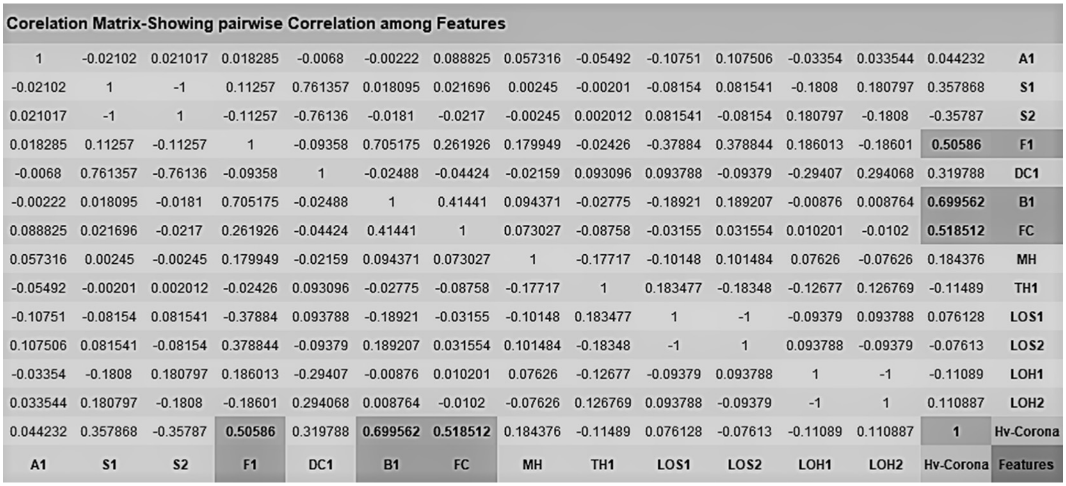


FIGURE 20.2 Machine learning (ML) process flowchart.

A correlation matrix, which is a tool for the feature selection process, is table used to define correlation coefficients among variables or features. Every cell in the matrix defines a correlation between two variables. It is used to summarize a large dataset and also to identify the most highly correlated features (shown in gray entries in second last row and column in Fig. 20.3) in the given data [35–37,40].

The correlation coefficient's value near 1 signifies that features participating in correlation are highly correlated to each other; on the other hand, the correlation coefficient's value near 0 signifies that features are less correlated to each other. Generally, correlation could be of two types: positive and negative. A positive correlation states that an increase or decrease in one feature's value results in an increase or



Corelation Matrix-Showing pairwise Correlation among Features

1	-0.02102	0.021017	0.018285	-0.0068	-0.00222	0.088825	0.057316	-0.05492	-0.10751	0.107506	-0.03354	0.033544	0.044232	**A1**
-0.02102	1	-1	0.11257	0.761357	0.018095	0.021696	0.00245	-0.00201	-0.08154	0.081541	-0.1808	0.180797	0.357868	**S1**
0.021017	-1	1	-0.11257	-0.76136	-0.0181	-0.0217	-0.00245	0.002012	0.081541	-0.08154	0.180797	-0.1808	-0.35787	**S2**
0.018285	0.11257	-0.11257	1	-0.09358	0.705175	0.261926	0.179949	-0.02426	-0.37884	0.378844	0.186013	-0.18601	**0.50586**	**F1**
-0.0068	0.761357	-0.76136	-0.09358	1	-0.02488	-0.04424	-0.02159	0.093096	0.093788	-0.09379	-0.29407	0.294068	0.319788	**DC1**
-0.00222	0.018095	-0.0181	0.705175	-0.02488	1	0.41441	0.094371	-0.02775	-0.18921	0.189207	-0.00876	0.008764	**0.699562**	**B1**
0.088825	0.021696	-0.0217	0.261926	-0.04424	0.41441	1	0.073027	-0.08758	-0.03155	0.031554	0.010201	-0.0102	**0.518512**	**FC**
0.057316	0.00245	-0.00245	0.179949	-0.02159	0.094371	0.073027	1	-0.17717	-0.10148	0.101484	0.07626	-0.07626	0.184376	**MH**
-0.05492	-0.00201	0.002012	-0.02426	0.093096	-0.02775	-0.08758	-0.17717	1	0.183477	-0.18348	-0.12677	0.126769	-0.11489	**TH1**
-0.10751	-0.08154	0.081541	-0.37884	0.093788	-0.18921	-0.03155	-0.10148	0.183477	1	-1	-0.09379	0.093788	0.076128	**LOS1**
0.107506	0.081541	-0.08154	0.378844	-0.09379	0.189207	0.031554	0.101484	-0.18348	-1	1	0.093788	-0.09379	-0.07613	**LOS2**
-0.03354	-0.1808	0.180797	0.186013	-0.29407	-0.00876	0.010201	0.07626	-0.12677	-0.09379	0.093788	1	-1	-0.11089	**LOH1**
0.033544	0.180797	-0.1808	-0.18601	0.294068	0.008764	-0.0102	-0.07626	0.126769	0.093788	-0.09379	-1	1	0.110887	**LOH2**
0.044232	0.357868	-0.35787	**0.50586**	0.319788	**0.699562**	**0.518512**	0.184376	-0.11489	0.076128	-0.07613	-0.11089	0.110887	**1**	**Hv-Corona**
A1	**S1**	**S2**	**F1**	**DC1**	**B1**	**FC**	**MH**	**TH1**	**LOS1**	**LOS2**	**LOH1**	**LOH2**	**Hv-Corona**	**Features**

FIGURE 20.3 Correlation matrix values.

decrease in the other feature's value; in contrast, a negative correlation has a reverse relation between the two features, so an increase in one feature's value results in the decreased value of the other feature.

Rows and columns in the correlation matrix represent each feature's name. Each cell in a table containing the correlation coefficient calculated between features corresponds to the respective row and column of that particular cell.

Fig. 20.4 shows another form of representation of a correlation matrix using a heat map. Heat maps are a popular way to visualize the interrelation between two or more variables or features, because it is easy for the human mind to distinguish between an attribute's ranks by visualizing color coding rather than checking and searching for the best value in a given list of numerical values, as shown in Fig. 20.3. One can easily identify and choose the most correlated feature using heat map visualization, in which the light-colored cell defines the most correlated features and the dark-colored cell defines the least correlated features.

Fig. 20.5 shows the prediction performance based on two classifiers, random forest and ETC. The ETC gives one wrong prediction (dark gray colored column) out of 14 data points and the random forest classifier (RFC) gives three wrong predictions out of 24 data points.

Fig. 20.6 compares two classifier outputs using line graphs: ETC and RFC. The figure shows 24 data points, on the basis of which the accuracy of these classifiers is described.

The ETC misclassified at point 16, whereas the RFC misclassified at three points: 2, 16, and 24. This means the ETC is more accurate than the RFC. This comparison is shown for synthetic data. thus, if real data are used, based on those data, training of the classifier is performed and then the classifier is tested for accuracy. The classifier that gives the best accuracy can be used for prediction.

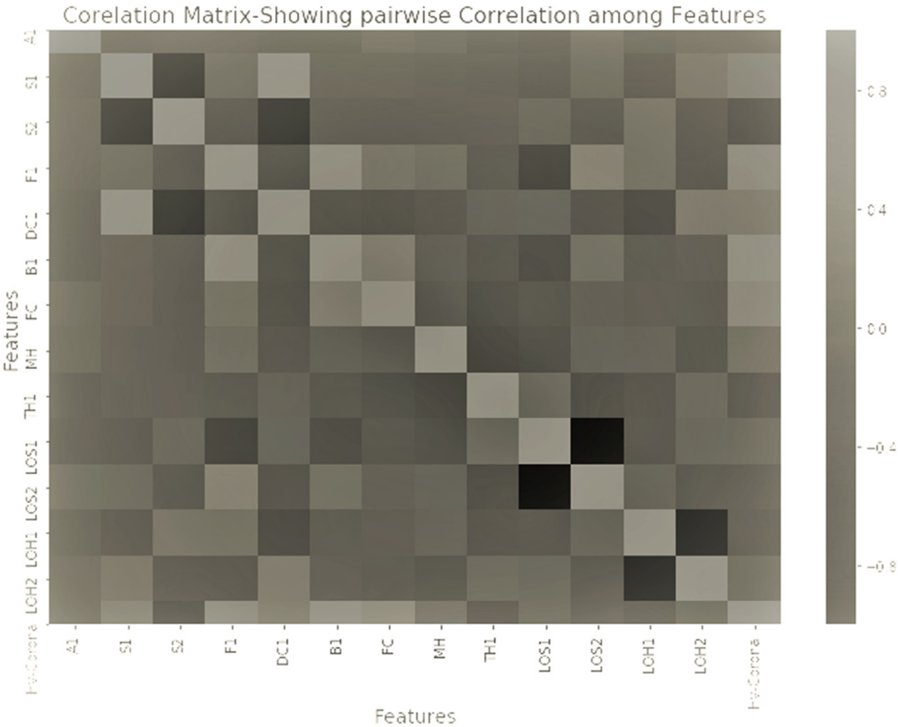
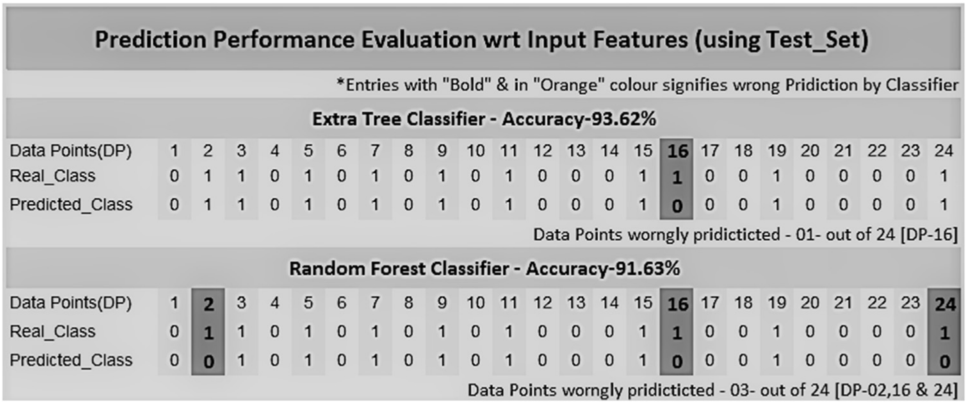


FIGURE 20.4 Representation of correlation matrix using heat map.



Prediction Performance Evaluation wrt Input Features (using Test_Set)

*Entries with "Bold" & in "Orange" colour signifies wrong Pridiction by Classifier

Extra Tree Classifier - Accuracy-93.62%

Data Points(DP)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	**16**	17	18	19	20	21	22	23	24
Real_Class	0	1	1	0	1	0	1	0	1	0	1	0	0	0	1	**1**	0	0	1	0	0	0	0	1
Predicted_Class	0	1	1	0	1	0	1	0	1	0	1	0	0	0	1	**0**	0	0	1	0	0	0	0	1

Data Points worngly pridictiected - 01- out of 24 [DP-16]

Random Forest Classifier - Accuracy-91.63%

Data Points(DP)	1	**2**	3	4	5	6	7	8	9	10	11	12	13	14	15	**16**	17	18	19	20	21	22	23	**24**
Real_Class	0	**1**	1	0	1	0	1	0	1	0	1	0	0	0	1	**1**	0	0	1	0	0	0	0	**1**
Predicted_Class	0	**0**	1	0	1	0	1	0	1	0	1	0	0	0	1	**0**	0	0	1	0	0	0	0	**0**

Data Points worngly pridictiected - 03- out of 24 [DP-02,16 & 24]

FIGURE 20.5 Prediction performance evaluation wrt input features.

7. Forecasting

For forecasting through ML, time series analysis may be used, which is an important part of ML. It is a univariate type of regression in which the target feature (dependent feature) is forecast using only one input feature (independent feature), which is time [41–43].

It is used to forecast future event values, and it has an important role for forecasting the existence of respiratory diseases such as COVID-19. Positive cases are increasing daily, so it is necessary to forecast whether the ratio by which the number is increasing is

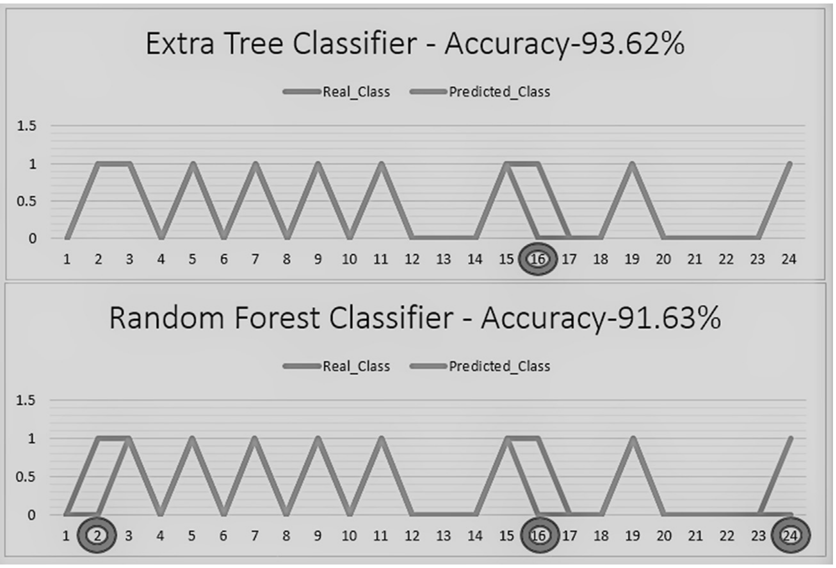


FIGURE 20.6 Comparison of extra tree classifier and random forest classifier.

continuing based on prior observations. It is helpful for the government, because based on the forecast, it can plan for resources to control the spread of disease and act for the future so that the growth rate of the infection decreases without affecting more people [30,32,35].

Forecasts depend completely on past trends, so forecast values cannot be guaranteed. However, this forecasted approximation of events may help authorities to assess forthcoming resource planning to compete with any pandemic situation such as COVID-19.

We used the most widely used forecasting method, called the ARIMA model for time series forecasting. ARIMA is used for time series data to predict future trends [41–50].

ARIMA is a form of univariate regression analysis that predicts future values based on differences between values rather than actual values. It combines three terms: (1) autoregression (AR), showing changing a variable that revert values on the basis of its prior values; (2) integrated (I), or replacing data values on the basis of the difference between data values and previous values; and (3) moving average (MA), a successive average taken on successive time frames of constant size of a time series previously available.

ARIMA uses a pdq forecasting equation in which the these parameters are defined as:

p is the number of observations that have lagged;

d defines the time (i.e., how many times the raw observations are different); and

q defines the size of the MA window.

In the AR model, the value of Yt depends on its own lagged value. In the MA model, the value of Yt depends on lagged forecast errors. Thus, the general equation of ARIMA is:

$$\mathbf{Yt = \alpha + \beta 1Yt - 1 + \beta 2Yt - 2 + \ldots + \beta pYt - p + \varepsilon 1 + \theta 1\varepsilon t - 1 + \theta 2\varepsilon t - 2 + \ldots + \theta q\varepsilon t - q}$$

where Yt is the target to be predicted, α is the constant value, $\beta1Yt - 1$ is the linear combination lags of Y that are taken up to p lags, and $\Theta1\varepsilon t - 1$ is the linear combination of lagged forecast error that is taken up to q lags [41–49].

The syntax of the ARIMA model is:

$$\mathbf{ARIMA}(<\mathbf{Dataset\_Name}>[\text{‘}\mathbf{Targetvalue\_Name}\text{’}],\mathbf{order}=(\mathbf{p},\mathbf{d},\mathbf{q}))$$

For example, ar_model = ARIMA(data[‘confirmed_COVID_cases’], order = (3,1,0)

We used the COVID-19 dataset (covid_19_india.csv) until Apr. 19, 2020 from Kaggle [34]; based on this dataset, we forecast confirmed cases in India and the top 10 states (with respect to COVID-19 infection cases), performed using the ARIMA method.

As good ML practice, dataset was pre-processed (only required features have been selected) and dataset has been split into two parts Training Set (30-01-2020–15-04-2020) and Test(Validation)_Set (15-04-2020–19-04-2020).

After that, the model was trained using the training set employing several pdq configurations of the ARIMA model and then cross-validated using the testing set. Results of the cross-validation are listed in Table 20.1.

In Table 20.1, the error rate (root mean square error [RMSE]) of the model for different states is shown using the ARIMA model. Entries in bold show the lowest RMSE for the state of a particular pdq configuration. The lowest value of RMSE is treated as the best configuration of ARIMA to forecast future values for a particular state. According to the values given in Table 20.1, we have selected two states, Telangana and West Bengal, and the total cases for India.

Fig. 20.7 shows the forecast for Telangana state based on data up to Apr. 19, 2020. *Dark gray dots* signify the actual training set (past real observation) upon which the model was trained and green (gray in printed version) signifies the actual testing set (partial data points from the dataset) for which the furcated value was validated (see overlapping area of *light gray dots*); pink (black in printed version) is the future forecast.

Table 20.1 Results.

ARIMA ⟶	**Parameter (p, d, q) configuration**		
State ↓	**5,1,0**	**3,1,0**	**1,1,0**
Maharashtra	320.31	284.49	128.82
Gujarat	454.54	394.30	379.00
Delhi	287.27	319.62	125.04
Rajasthan	48.29	75.30	78.25
Madhya Pradesh	92.33	107.07	116.64
Tamil Nadu	105.02	98.24	51.59
Uttar Pradesh	154.22	37.23	144.01
Telangana	12.66	17.97	99.49
Andhra Pradesh	13.59	14.28	29.93
West Bengal	10.90	3.38	21.28
All India	1299.87	758.60	253.60
	Error rate (root mean square error value) ↑		

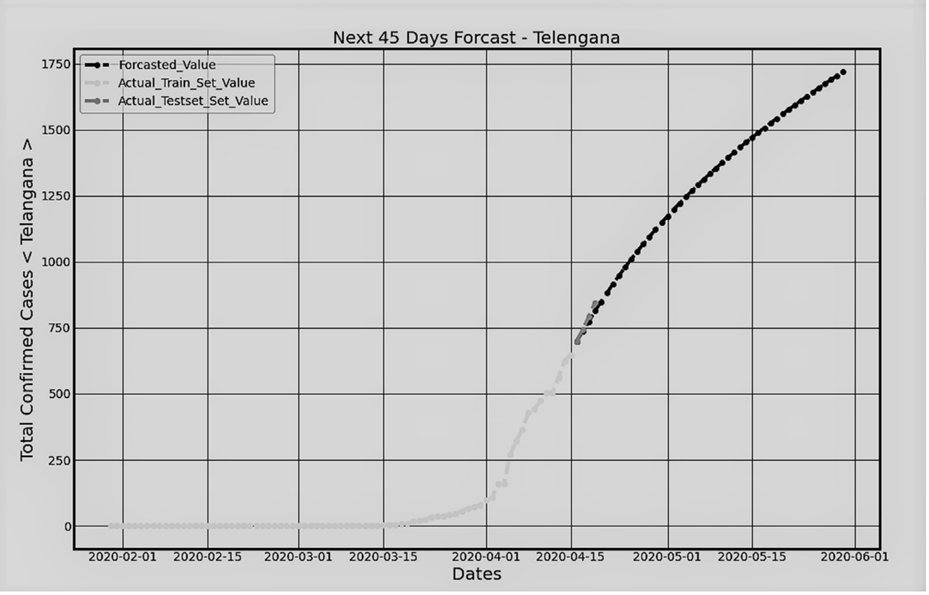


FIGURE 20.7 Forecast for Telangana.

As displayed in the graphs (Figs. 20.7–20.9) we have forecast future values with confirmed cases for Telangana, West Bengal, and India for the next 45 days using past data that was available until Apr. 19, 2020. This model can be used to forecast future values based on past values of COVID-19 and similar infectious diseases (if actual data are available for testing).

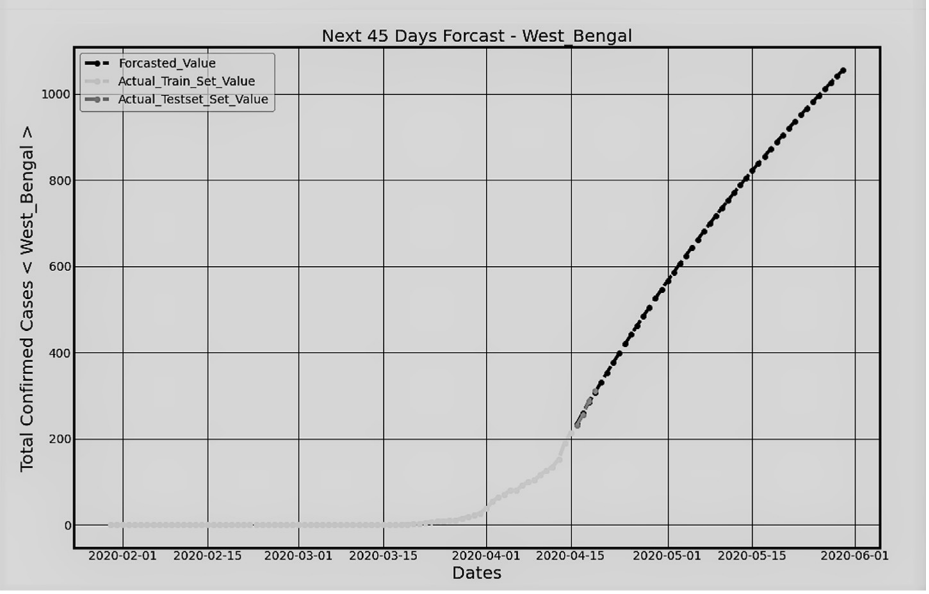


FIGURE 20.8 Forecast for West Bengal.

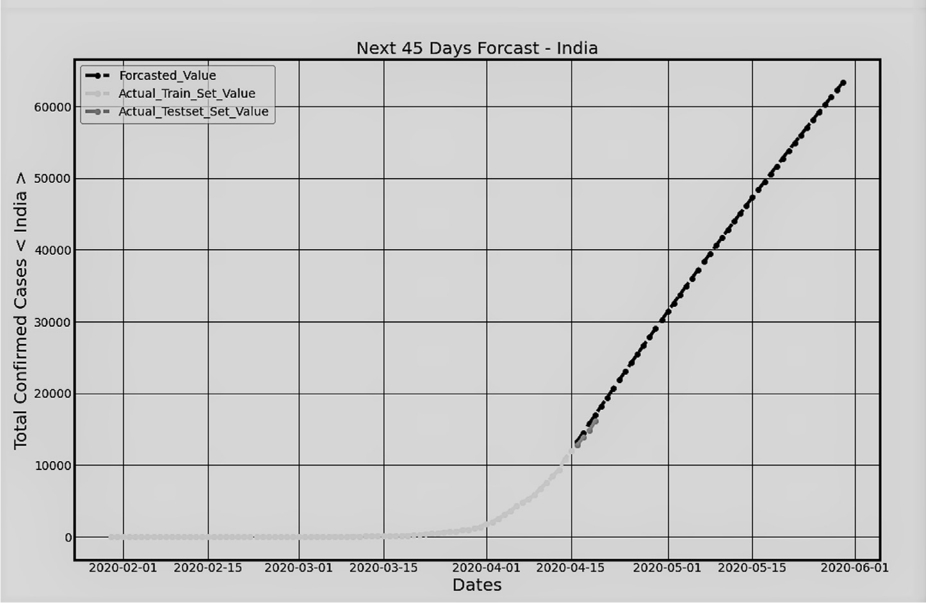


FIGURE 20.9 Forecast for India.

8. Conclusion and future work

The pandemic of COVID-19 has affected the entire globe. It has spread in more than 85 countries as of Apr. 2020. Scientists have made every effort to find solutions to it; according to claims by the United States and India, some vaccines have been made that are being trialed. The use of computers by scientists for early prediction has been widespread. A lot of research is taking place using ML to combat COVID-19. This chapter can be used by different researchers to learn how ML can be employed to forecast not only this situation but also other cases. The chapter specifically used the ARIMA method of time to forecast the stability and growth of COVID-19. Many countries have seen high totals of deaths owing to COVID-19. It is believed that the performance of the model can be improved or the model can give more accurate data if more datasets are available. The model gives results on the basis of data developed by information given by health agencies. Thus, forecasting may not be 100% accurate, but it can surely be used as a corrective measure. For future work further enhancement can be done by combining new factors and algorithms with ARIMA to get more accurate results.

References

[1] Confirmed cases of Covid 19. Available from: https://www.covid19india.org/. (Accessed 26 April 2020).

[2] David J Cennimo Discusses Coronavirus Disease 2019 (COVID 19). Available from: https://emedicine.medscape.com/article/2500114-overview. (Accessed 25 April 2020).

[3] J. Wang, Y. Liu, Y. Wei, J. Xia, T. Yu, X. Zhang, L. Zhang, Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study, Lancet 395 (10223) (2020) 507–513.

[4] Coronavirus Disease (COVID 19) Outbreak. Available from: http://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/novel-coronavirus-2019-ncov. (Accessed 25 April 2020).

[5] COVID 19 Article. Available from: https://www.newscientist.com/term/covid-19/. (Accessed 30 April 2020).

[6] Erica Hersh Discusses How Long Is the Incubation Period for the Coronavirus?. Available from: https://www.healthline.com/health/coronavirus-incubation-period#incubation-period. (Accessed 25 April 2020).

[7] Symptoms of Coronavirus. Available from: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html. (Accessed 24 April 2020).

[8] Anulekha Ray Discusses ABOUT Coronavirus: India's biggest Concerns are COVID 19 Patients with No Symptoms. Available from: https://www.livemint.com/news/india/coronavirus-india-s-biggest-concerns-are-covid-19-patients-with-no-symptoms-11587533159071.html. (Accessed 26 April 2020).

[9] Teena Thacker Discusses About No Symptoms in 80% of COVID Cases Raises Concern. Available from: https://economictimes.indiatimes.com/industry/healthcare/biotech/healthcare/no-symptoms-in-80-of-covid-cases-raise-concerns/articleshow/75260387.cms?from=mdr. (Accessed 26 April 2020).

[10] Praveen Duddu Discusses About COVID 19 Coronavirus: Top Ten Most Affected Countries. Available from: https://www.pharmaceutical-technology.com/features/covid-19-coronavirus-top-ten-most-affected-countries/.

[11] Covid 19 Cases in China. Available from: https://www.worldometers.info/coronavirus/country/china/. (Accesses 26 April 2020).

[12] V. Wang, Coronavirus Epidemic Keeps Growing, But Spread in China Slows. New York Times. https://www.nytimes.com/2020/02/18/world/asia/china-coronavirus-cases.html?referringSource=articleShare. (Accessed 26 April 2020).

[13] Covid 19 cases in Italy. Available from: https://www.worldometers.info/coronavirus/country/italy/. (Accessed 26 April 2020).

[14] Covid 19 cases in Spain. Available from: https://www.worldometers.info/coronavirus/country/Spain/. (Accessed 26 April 2020).

[15] Covid 19 cases in US. Available from: https://www.worldometers.info/coronavirus/country/US/. (Accesses 26 April 2020).

[16] Covid 19 cases in Germany. Available from: https://www.worldometers.info/coronavirus/country/Germany/. (Accesses 26 April 2020).

[17] Covid 19 cases in France. Available from: https://www.worldometers.info/coronavirus/country/France/. (Accesses 26 April 2020).

[18] Covid 19 cases in Iran. Available from: https://www.worldometers.info/coronavirus/country/Iran/. (Accesses 26 April 2020).

[19] Covid 19 cases in Turkey. Available from: https://www.worldometers.info/coronavirus/country/Turkey/. (Accesses 26 April 2020).

[20] Gavin Edwards discusses about Machine Learning: An Introduction. Available from: https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0. (Accessed 27 April 2020).

[21] D. Gupta, A. Julka, S. Jain, T. Aggarwal, A. Khanna, N. Arunkumar, V.H.C. De Albuquerque, Optimized cuttlefish algorithm for diagnosis of Parkinson's disease, Cognit. Syst. Res. 52 (2018) 36–48.

[22] R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA, 2014.

[23] I. Turaiki, M. Alshahrani, T. Almutairi, Building predictive models for MERS-CoV infections using data mining techniques, J. Infect. Public Health 09 (2016) 744—748.

[24] V. Chouhan, S.K. Singh, A. Khamparia, D. Gupta, P. Tiwari, C. Moreira, R. Damaševičius, V.H.C. De Albuquerque, A Novel transfer learning based approach for pneumonia detection in chest X-ray images, Appl. Sci. 10 (2020) 559.

[25] S. Sreeja, L. Bhavya, S. Swamynath, R. Dhanuja, Chest x-ray pneumonia prediction using machine learning algorithms, Int. J. Res. Appl. Sci. Eng. Technol. 07 (04) (2019) 3227—3230.

[26] U. Kose, G.E. Guraksin, O. Deperlioglu, Diabetes Determination via Vortex Optimization Algorithm Based Support Vector Machines: Medical Technologies National Conference, 2015, pp. 1—4.

[27] S. Sharmila, C. Dharuman, P. Venkatesan, Disease classification using machine learning algorithms — a comparative study, Int. J. Pure Appl. Math. 114 (06) (2017) 1—10.

[28] S.S. Shirsath, Disease prediction using machine learning over big data, Int. J. Innov. Res. Sci. 07 (06) (2018) 6752—6757.

[29] O. Er, N. Yumusak, F. Temurtas, Chest diseases diagnosis using artificial neural networks, Expert Syst. Appl. 37 (12) (2010) 7648—7655.

[30] Z.F. Yang, Z.Q. Zeng, K. Wang, et al., Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions, J. Thorac. Dis. 12 (3) (2020) 165—174.

[31] M.A. Müller, V.M. Corman, J. Jores, B. Meyer, M. Younan, A. Liljander, B.J. Bosch, E. Lattwein, M. Hilali, B.E. Musa, S. Bornstein, MERS coronavirus neutralizing antibodies in camels, Eastern Africa, 1983—1997, Emerg. Infect. Dis. 20 (12) (2014) 2093.

[32] A.A. El-Solh, C.B. Hsiao, S. Goodnough, J. Serghani, B.J. Grant, Predicting active pulmonary tuberculosis using an artificial neural network, Chest 116 (04) (1999) 968—973.

[33] U.K. Tiwari, Role of machine learning to predict the outbreak of covid-19 in India, J. Xi'an Univ. Archit. Technol. 12 (4) (2020) 2663—2669.

[34] S. Makridakis, A. Wakefield, R. Kirkham, Predicting medical risks and appreciating uncertainty, Foresight Int. J. Appl. Forecast. 52 (2019) 28—35.

[35] S. Tuli, S. Tuli, R. Tuli, S.S. Gill, Predicting the Growth and Trend of COVID-19 Pandemic Using Machine Learning and Cloud Computing, Internet of Things, 2020.

[36] N.S. Punn, S.K. Sonbhadra, S. Agarwal, COVID-19 Epidemic Analysis Using Machine Learning and Deep Learning Algorithms, MedRxiv, 2020.

[37] L. Jia, K. Li, Y. Jiang, X. Guo, Prediction and Analysis of Coronavirus Disease 2019. arXiv Preprint, 2020, p. 05447, arXiv:2003.

[38] G. Kalipe, V. Gautham, R.K. Behera, Predicting Malarial Outbreak Using Machine Learning and Deep Learning Approach: A Review and analysis; International Conference on Information Technology (ICIT), IEEE, 2018, pp. 33—38.

[39] O. Er, N. Yumusak, F. Temurtas, A.C. Tanrikulu, A comparative study on chronic obstructive pulmonary and pneumonia diseases diagnosis using neural networks and artificial immune system, J. Med. Syst. 33 (06) (2009) 485—492.

[40] S. Khobragade, A. Tiwari, C. Patil, V. Narke, Automatic Detection of Major Lung Diseases Using Chest Radiographs and Classification by Feed-Forward Artificial Neural Network: IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems, IEEE, 2016, pp. 1—5.

[41] U. Kose, S. Grima, E. Özen, H. Boz, J. Spiteri, E. Thalassinos, Using artificial intelligence techniques for economic time series prediction, in: Contemporary Issues in Behavioral Finance 101, 2019, pp. 13—28.

[42] U. Kose, An ant-lion optimizer-trained artificial neural network system for chaotic electroencephalogram (EEG) prediction, Multidisciplinary Digital Publishing Institute(MDPI). 08 (09) (2018) 1613.

[43] Sudalai Rajkumar discusses about datasets. Available from: https://www.kaggle.com/sudalairajkumar/covid19-in-india. (Accessed 28 April 2020).

[44] M.J. Kane, N. Price, M. Scotch, et al., Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks, BMC Bioinf. 15 (01) (2014) 1−9.

[45] Q. Liu, X. Liu, B. Jiang, W. Yang, Forcasting incidence of hemorrhagic fever with renal syndrome in China using arima model, BMC Infect. Dis. 11 (218) (2011).

[46] A.E. Hoerl, R.W. Kannard, K.F. Baldwin, Ridge regression − some simulations, Commun. Stat. 04 (02) (1975) 105−123.

[47] R.J. Hyndman, A.B. Koehler, R.D. Snyder, S. Grose, A state space framework for automatic forecasting using exponential smoothing methods, Int. J. Forecast. 18 (03) (2002) 439−454.

[48] Sanjay Sharma discusses about How predictive models can aid in the battle against COVID-19. Available from: https://home.kpmg/in/en/home/insights/2020/04/how-predictive-models-can-aid-in-the-battle-against-covid-19.html. (Accessed 30 April 2020).

[49] M.P. Bohdan, Machine-Learning Models for Sales Time Series Forecast: MDPI, vol. 04, 2019 (15).

[50] K.C. Santosh, AI-driven tools for coronavirus outbreak: need of active learning and cross-population train/test models on multitudinal/multimodal data, J. Med. Syst. 44 (2020) 93.