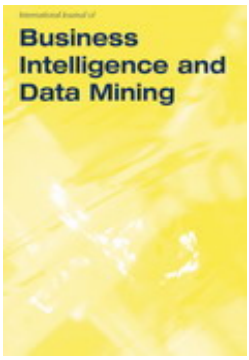


International Journal of Business Intelligence and Data Mining

ISSN online: 1743-8195 - ISSN print: 1743-8187
https://www.inderscience.com/ijbidm

Machine learning approach for data analysis and predicting coronavirus using COVID-19 India dataset

Soni Singh, K.R. Ramkumar, Ashima Kukkar

DOI: 10.1504/IJBIDM.2024.10049479

Article History:
Received: 18 January 2022
Accepted: 30 May 2022
Published online: 01 December 2023

Machine learning approach for data analysis and predicting coronavirus using COVID-19 India dataset

Soni Singh*, K.R. Ramkumar and Ashima Kukkar

Department of Computer Science and Engineering,
Chitkara Institute of Engineering and Technology,
Chitkara University, Punjab, India
Email: sonisingh0107@gmail.com
Email: k.ramkumar@chitkara.edu.in
Email: ashima@chitkara.edu.in
*Corresponding author

Abstract: According to the World Health Organisation (WHO), the COVID-19 virus would infect 83,558,756 persons worldwide in 2020, resulting in 646,949 deaths. In this research, we aim to find the link between the time series data and current circumstances to predict the future outbreak and try to figure out which technique is best for modelling for accurate predictions. The performance of different machine learning (ML) models such as sigmoid function, Facebook (FB) prophet model, seasonal auto-regressive integrated moving average with eXogenous factors (SARIMAX) model, support vector machine (SVM) learning model, linear regression (LR) model, and polynomial regression (PR) model are analysed along with their error rate. A comparison is also done to evaluate a best-suited model for prediction based on different categorisation approaches on the WHO authenticated dataset of India. The result states that the PR model shows the best performance with time-series data of COVID-19 whereas the sigmoid model has the consistently smallest prediction error rates for tracking the dynamics of incidents. In contrast, the PR model provided the most realistic prediction to identify a plateau point in the incident's growth curve.

Keywords: COVID-19; pandemics; analysis on India; machine learning; prediction; comparison; support vector machine; SVM.

Reference to this paper should be made as follows: Singh, S., Ramkumar, K.R. and Kukkar, A. (2024) 'Machine learning approach for data analysis and predicting coronavirus using COVID-19 India dataset', *Int. J. Business Intelligence and Data Mining*, Vol. 24, No. 1, pp.47–73.

Biographical notes: Soni Singh is a PhD candidate in the Department of Computer Science and Engineering, Chitkara University, Punjab. She has been doing research in artificial intelligence, machine learning, and deep learning since 2017. Her current research is concerned with the prediction of pandemic outbreaks using machine learning and deep learning techniques. She has worked on the Pregaura project for the women's healthcare system.

K.R. Ramkumar is an Associate Professor in the Department of Computer Science and Engineering, Chitkara University, Punjab. His areas of expertise are network security, key management and relational database management systems with advancements. His research includes solving the routing issues and dealing with security and node failure apprehensions of wireless sensor

networks. Much of his work has been on improving the understanding, design, and performance analysis of different routing and security algorithms of wireless sensor networks. He also is working with the extensible markup language (XML) and resolving the data integrity and consistency issues in web communications. During his rest time, he does research on green, free and sustainable energy.

Ashima Kukkar is an Assistant Professor in the Department of Computer Science and Engineering, Chitkara University, Punjab. She has been doing research in text classification, data mining, natural language processing, text summarisation, swarm intelligence, recommendation systems, and deep learning since 2015. Her current research is concerned with natural language processing, deep learning, machine learning, text classification, recommendation systems, meta-heuristic algorithms, pattern recognition, and data mining techniques.

1 Introduction

The novel Coronavirus (COVID-19) is a transferable disease that was first detected in China in December 2019 and then spread throughout the world. On 11 March 2020, World Health Organisation (WHO) declared Novel Coronavirus disease as a pandemic and reiterated the call for countries to take immediate actions and scale up the response to treat, detect and reduce transmission to save people's lives (Muhilthini et al., 2018). According to a WHO report in the year 2020 only, a total of 83,558,756 people got affected by the COVID-19 virus with 646,949 deaths. Worldwide, India is the second most affected country with 31,174,322 COVID-19 confirmed cases as of July 20, 2021, with 414,482 deaths and 30,353,710 recoveries but due to the virus's different variants, there may be chances of infection again in those recovered cases (Punn et al., 2020; Yadav et al., 2020; Liu et al., 2021; Ardabili et al., 2020). The virus has spread widely, and the number of cases is rising daily as governments work to slow its spread. India has moved quickly, implementing a proactive, nationwide, lockdown, to flatten the curve and use the time to plan and resource responses adequately (Asher, 2018). Figure 1 shows the graphical plot of total cases, total deaths, actual cases and total closed cases (recovery rate vs. death rate) in India from January 22, 2020, to May 29, 2021.

The government of India has taken several steps to control the spread of the virus such as the closure of schools, testing, quarantine and vaccination. There are many other actions taken by the government like complete lockdown across India, lockdown 1, lockdown 2, unlock 1 with restrictions and unlock 2 with some restrictions, and took many steps but still, the spread rate of the virus is very high. Further, the nation was divided into three zones red zones (hotspots), green zones (not hotspots) and yellow zones (suspected areas) (Abirami and Chitra, 2020). Based on the spread rate of COVID-19 in India the pandemic is divided into first and second waves. In the first wave, the infection rate and death rate were very low but in the second wave, the infection and death rate are very high. During the first wave, the government also faces a huge economic fall down (Wu et al., 2020; Tian et al., 2020; Khan et al., 2021). Tourism, retail sector, hospitality and many sectors face huge losses aimed COVID-19 pandemic. Thousands of people lost their jobs due to this crisis. Vaccination of people is also started in India. Many people have already taken their first dose of vaccination. But still the virus

mutates each time and affects different people in different ways, with most infected people having mild to moderate illness and recovering without hospitalisation and some with severe infections which cause death. The outbreak differs from other recent outbreaks, which challenge the standard model machine learning (ML) models to provide accurate results (Nazir and Khan, 2021). In addition to the many known and unknown variables involved in the prevalence, complexity, and control of population-wide behaviour in different geopolitical regions, the strategies had dramatically increased model uncertainty. As a result, standard epidemiological models pose new challenges to produce more definitive results. To overcome this challenge, many novel models have come to light that suggests diver's expropriation to modelling (e.g., quarantines, lockdown, social distancing and vaccination, etc.) to overcome this pandemic (Wannier et al., 2019; Rongali and Yalavarthi, 2017; Chen, 2015; Josephine et al., 2021).

The main purpose of this research work is to propose the development of a prognostic tool for COVID-19 prediction with improved accuracy. There have been a huge amount of data and datasets available on the internet or from external sources and the COVID-19 dataset which has been used in this work is one of the most widely used datasets in many types of research and it is collected by the WHO (Françoise et al., 2021). This research work represents a comprehensive study done on the COVID datasets through the data visualisation process by using a model like sigmoid function, FB prophet, SRIMAX, SVM, LR and PR. The proposed model is completely fitted into the dataset which contains the total number of confirmed cases, deaths and the number of cured cases in India. Using this dataset, we analyse the current situation in India and also, and we predicted a number of confirmed, death and cured cases in India. It also helps in providing a solution for a future aspect to manage the COVID-19 pandemic. The comparison of the ML algorithm shows which define logical and RMSE score to conclude which algorithm is best suited for the accurate and prominent result.

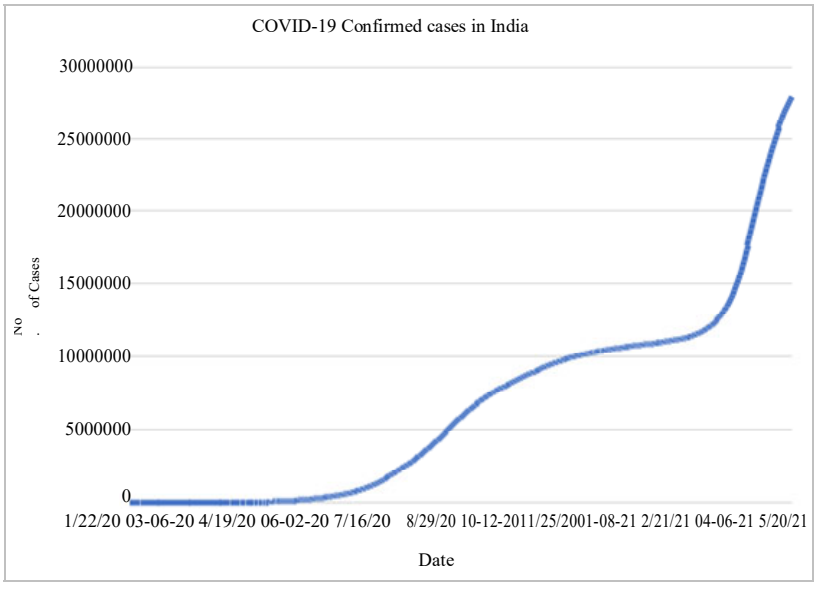
The pandemic's spread rate is still unpredictable, so, in this paper, the basic prediction model is defined as shown in Figure 2. Data collection, data cleaning, extraction and validation, ML predictor, and prediction output are the four primary components of the basic prediction model. The entire epidemic prediction process uses ML. These ML prediction models strive to create a model that performs well and reduces training error (Ardabili et al., 2020). The model error must be minimised since it shows the relationship between the actual value of the estimated learning parameter and the model forecast. However, the quality of training is determined by the available training samples as well as learning parameters. The key to ML algorithms that are learned from past training samples which helps in the prediction of the transmission rate of pandemic diseases. The modelling strategy considered three classes of mixed population.

The detailed contribution is described below:

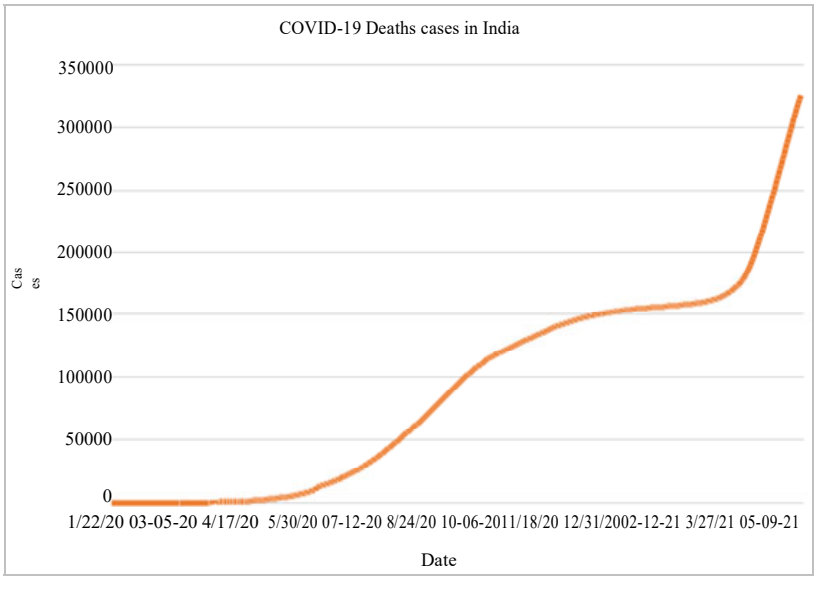
To analyse and calculate the growth factor of COVID-19 India cases for prediction.

- An existing prediction model is used to utilise ML techniques for time series predicting of pandemic disease outbreaks.
- In the prediction model, different techniques such as sigmoid, FB prophet, SRIMAX, SVM, LR, and PR are used to optimise the learning parameters so that the prediction error is minimised.

Figure 1 COVID-19 India information report by WHO from January 22, 2020, to May 29, 2021, (a) total confirmed cases of COVID-19 in India, (b) total COVID-19 deaths were reported in India, (c) total COVID-19 cured cases in India (see online version for colours)

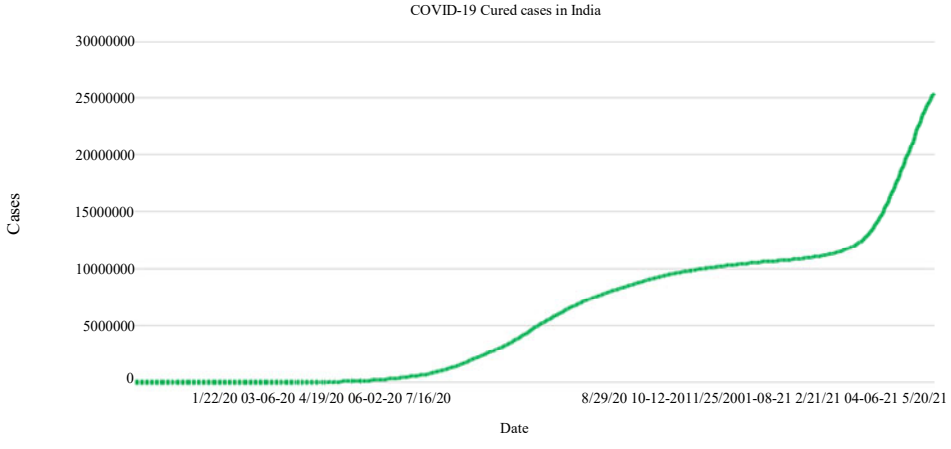


(a)



(b)

Figure 1 COVID-19 India information report by WHO from January 22, 2020, to May 29, 2021, (a) total confirmed cases of COVID-19 in India, (b) total COVID-19 deaths were reported in India, (c) total COVID-19 cured cases in India (continued) (see online version for colours)

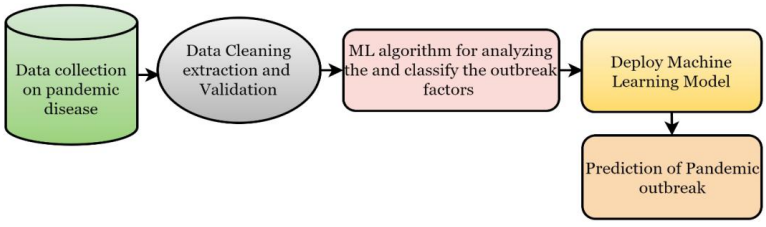


(c)

The result of the models is measured using the root mean square error (RMSE) of the model.

The design prediction model shows improved performance results when compared with existing ML models.

Figure 2 Stages involve in designing of ML model (see online version for colours)



The rest of the paper is organised as follows. The section review of related literature describes the previous work done for COVID-19 prediction. The next section materials and methods describe the methods used to forecast COVID-19 mortality in India. The approach for COVID-19 confirmed the death, and recovered cases prediction is detailed in section methodology. The section result and comparative analysis examine the most suitable prediction model. The last section conclusion and future work outlines this work as well as potential future work.

2 Review of related literature

In this research the author proposed a prediction model that uses gradient boosting regression (GBR) and mean square error (MSE) for estimating the performance rate of

the model (Muhilthini et al., 2018). They collected data set for dengue disease which gives information about patients affected by dengue. The dengue dataset shows important aspects like temperature and rainfall, etc. which is the main cause of dengue disease. The proposed model GBR model can work with different types of data for predicting the impact of the disease. MSE is used to estimate the units and also used to measure the bitterness of the model that helps in analysing and predicting the dengue disease in future.

Punn et al. (2020) proposed the use of ML and deep learning techniques to analyse pandemics with the help of collected data from the online portal Johns Hopkins dashboard. After experimenting, the result described that the polynomial regression (PR) provides a minimum RMSE as compared to the other approaches in forecasting data transmission for COVID-19, if the spreading of the virus follows the predicted trend of the PR model, then the spread rate of the virus would be high and it leads to huge life loss.

In this paper the researcher designs a model to fight against a coronavirus disease (Yadav et al., 2020). The main focus of the author is generating and comparing the outcomes or results of different regression models on the available data from the trusted source on online portals. In this work they set five objectives based on the current situation of the COVID-19 pandemic they are:

1 Analyse and predict the spread rate of this deadly virus worldwide.

2 Identify the type of risk assessment and growth rate of coronavirus across the world.

3 Provide solutions and predictions over this pandemic.

4 Measure the transmission rate of this virus.

5 Comparing and correlating the COVID-19 pandemic with the weather.

Tian et al. (2020) proposed that the novel coronavirus COVID-19 has caused the outbreak of pneumonia first detected in Wuhan, Hubei province, china and now its spread all over the world. To date, there are no clinically approved drugs or vaccines are available to control this viral pandemic. He describes some epidemic and etiological characteristics of COVID-19 and also discusses some biological features of this pandemic including tropism and receptor usage, at the end, he discusses the prevention and treatment and transmission of COVID-19.

Khan et al. (2021) proposed to regulate a literature study to understand the COVID-19 pandemic and the role of ML to fight against this deadly virus. They also conduct a study on the epidemic situation, safety kit, diagnosis and other important aspects that help in controlling the spread of coronavirus. They study 128 high-quality articles and reviewed their important factor to understand the behaviour of this virus and how ML plays a big role to fight against this COVID-19 pandemic. The study shows important factors that are data types and other available aspects that help in providing medical assistance and other help in research.

Nazir and Khan (2021) proposed a prediction model using a ML algorithm for predicting the spread rate of coronavirus in countries. They perform their research on cases in Morocco country. After applying a ML algorithm, the result describes that the prediction rate of the model is very high and accurate. The score of the model is high they apply the same model to COVID-19 cases in China and the prediction is more relevant.

Thus, the model uses data from any country to predict measuring the spread rate of this COVID-19 pandemic.

Wannier et al. (2019) proposed a ML model to predict the behaviour, spread rate of the COVID-19 virus, spread period, peak point, size and active cases extend in India. The data set used for this research is from an online website they used this data set on autoregressive integrated moving average (ARIMA) model. To measure the performance of the model they use three regression models that are neural network (NN), support vector regression (SVR) and linear regression (LR) model to make a simple mean aggregated method. Using these models, they make the comparison that which model had a better prediction rate.

MacIntyre et al. (2021) proposed to develop an AI-based model for the improvement of critical care for COVID-19 patients. The author performs the research with the help of available data in PubMed, Web of science and other data. Based on the available data studies a three-stage model of input, process and output was created. This model proposed the AI application in the intensive care unit. The ML techniques help in the fight against COVID-19 if we use an AI-based expert system then it would double up the management of the patient in a more efficient way, especially for those who were in COVID-19 ICU.

Gupta and Pal (2020) proposed the use of available latest information on worldwide AI for COVID-19 which helps in analysing the many possible model and applications to tackle the disease. He defined seven significant applications of AI for COVID-19 which help make an action plan and fight against this deadly virus by analysing all available previous data. This application also helps in better understanding disease and the development of the vaccine.

Kelly et al. (2019) proposed that the data from an opinion poll define that the fear of COVID-19 in people is increasing day by day due to the spread nature of the outbreak in more countries which also impacts the physiological nature of people.

Goldstein et al. (2011) proposed a model that uses a combination of supervised machine learning and digital signal processing (MLDSP) that is used for genome analysis using an augmented decision tree of the ML approach, and a Spearman's rank correlation coefficient that defines the checking of result. These tools are used to analyse a large dataset of over 5,000 unique viral genomic sequences, totalling 61.8 million bp, including the 29 COVID-19 virus sequences available on January 27, 2020.

Dhamodharavadhani et al. (2020) used a model such as probabilistic neural network (PNN), radial basis function neural network (RBFNN), and generalised regression neural network (GRNN) are used to create the COVID-19 mortality rate prediction (MRP) model for India. They used two datasets, D1 and D2. The performance of these models is measured using the RMSE and 'R', a correlation value between the actual and projected values. To increase prediction accuracy, the new hybrid models were created by integrating SNN models and the nonlinear autoregressive neural network (NAR-NN).

Iwendi et al. (2020) proposed a fine-tuned random forest model that was enhanced by the AdaBoost method. The model predicts the severity of the case and the potential outcome, recovery, or death based on the COVID-19 patient's geographical, travel, health, and demographic data. On the dataset used, the model has a 94% accuracy and an F1 score of 0.86. The data analysis demonstrates a positive association between patients' gender and fatalities, as well as the fact that the majority of patients are between the ages of 20 and 70.

3 Materials and methods

3.1 Data collection

The data set of COVID-19 is collected from the official repository of the WHO organisation (https://covid19.who.int/info/). The CSV file format is used as an input file. The data consist of the daily case reported of time series summary tables. The data contains attributes as date wise, confirm, death and recovered cases of the COVID-19 from January 22, 2020, to May 29 India. The dataset used for the training and testing process must be labelled dataset. Where we have used 70% data for the training process and 30% data is used for the testing process.

3.2 Data analysis

Data visualisation is the graphical representation of data which always helps to understand the raw data. For large amounts of data sets, data visualisation is essential to extract valuable information for developing any ML model. In this research, we are going to understand the growth factor of the number of confirmed cases, recovery and total deaths of COVID-19 in India (Xu et al., 2020). The data set here was used by us to study the COVID-19 in India state wise dataset. The main objective of using this data set is to measure the state-wise impact of the COVID-19 pandemic in India. It helps in analysing the current status of the COVID-19 virus in India. The dataset contains some parameters that help predict the COVID-19 virus in the coming week. The COVID-19 India data has nine columns and 17,786 rows. Date, state/UT, cures, deaths, and confirmed and active cases from January 22, 2020, to May 29, 2021, are the six most influential features that help predict COVID-19 (MacIntyre et al., 2021). All visualisations are done in Python by importing various libraries such as seaborn, matplotlib, subplots and plotly. COVID-19 India data is collected from various data sources. The collected from the WHO authenticated website. Data here we have used state-wise time series data of COVID-19 confirmed cases, deaths and recoveries across India. Using visualisation, we created several graphs for the total number of confirmed, death, recovered cases, recovery rate and mortality rate in India from January 2020 to May 2021 to understand India's position in this pandemic (Hojeong and Songhee, 2020).

Here in Figure 3 describe a state-wise analysis of COVID-19 cases in India. The list shows that Maharashtra tops the list. As we can see in the figure that the recovery rate is high in Punjab state as compared to other states. Similarly, West Bengal and Gujarat are showing high death rates (2020).

Figure 4 shows the graph for the number of active cases all over the state in India from Jan 2020 to May 2021. As we can see that Maharashtra state has the highest number of COVID-19 active cases and Andaman and Nicobar Islands state has the lowest number of COVID-19active cases.

Equations (1) and (2) have defined the recovery rate and death rate of confirmed cases in India from March 2020 to May 2021. The mortality rate and the recovery rate are calculated using the following formula

$$\text{Mortality rate} = (\text{Total Number of Death Cases}/\text{Total Number of Confirmed Cases}) \times 100 \quad (1)$$

$$\text{Recovery Rate} = (\text{Total Number of Recovered Cases}/\text{Total Number of Confirmed Cases}) \times 100 \quad (2)$$

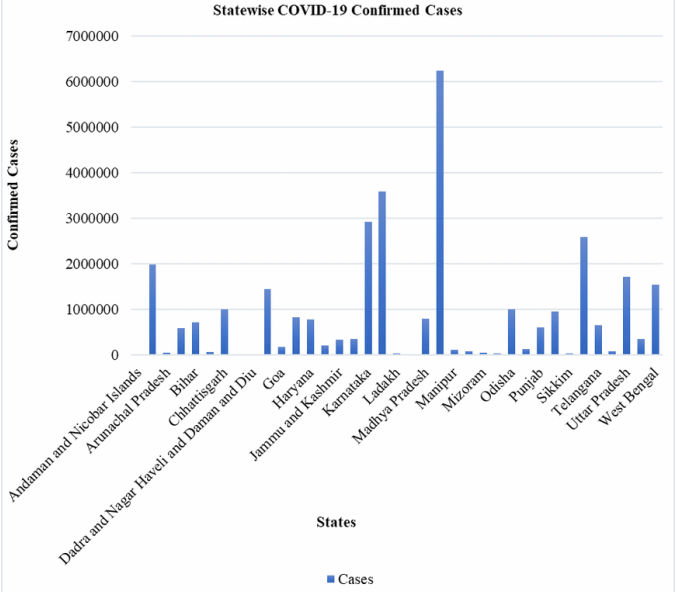
When the number of confirmed cases is lesser the rate of recovery is high. Initially, the increased mortality rate and the low recovery rate are a state of panic for India. An increase in the recovery rate is a good sign for India.

Figure 3 State-wise analysis of COVID-19 cases of confirmed, death, cured cases, recovery rate and mortality rate in India (see online version for colours)

```
statewise['Mortality Rate'] = statewise['Deaths']*100 /statewise['Confirmed']
statewise = statewise.sort_values(by='Confirmed', ascending= False)
statewise.style.background_gradient(axis=0)
```

State/UnionTerritory	Confirmed	Cured	Deaths	Recovery Rate	Mortality Rate
Maharashtra	6113335	5872268	123531	96.056702	2.020681
Kerala	2996094	2877557	13960	96.043615	0.465940
Karnataka	2859595	2784030	35526	97.357493	1.242344
Tamil Nadu	2503481	2435872	33132	97.299400	1.323437
Andhra Pradesh	1908065	1861937	12898	97.582472	0.675973
Uttar Pradesh	1706818	1682130	22656	98.553566	1.327382
West Bengal	1507241	1472132	17834	97.670645	1.183222
Delhi	1434687	1408853	25001	98.199328	1.742610
Chhattisgarh	996359	977893	13462	98.146652	1.351119
Rajasthan	952836	942882	8942	98.955329	0.938462
Odisha	927186	897362	4299	96.783385	0.463661
Gujarat	823964	811699	10072	98.511464	1.222384
Madhya Pradesh	790042	780578	9017	98.802089	1.141332
Haryana	769030	758442	9506	98.623201	1.236103
Bihar	722746	711913	9612	98.501133	1.329928

Figure 4 State-wise analysis of COVID-19 active cases in India from January 2020 to May 2021 (see online version for colours)



As we can see in Figure 5, the growth factor of the number of active cases in Lockdown 4.0 is slightly lower. The Prime Minister of India announced the lockdown on 24 March 2020 (Tian et al., 2020; Miranda et al., 2019). The above graph shows the growth factor of active cases during the lockdown and without lockdown. No lockdown period from 2020-01-30 to 2020-03-24, Announcement of Lockdown 1.0 from 2020-03-24 to 2020-07-15, Declaration of Lockdown 2.0 from 2020-07-15 to 2020-11-04 Announcement, Lockdown 3.0 Announcement from 2020-11-04 to 2021-02-19, Unlock 1.0 Announcement from 2020-06-01 to 2020-06-30 and Unlock 2.0 Announcement from 2020-07-01 to present.

Figure 5 The COVID-19 growth rate of active cases in India from January 2020 to May 2021 during lockdown and unlock (see online version for colours)

Table 1 COVID-19 active cases growth rate during lockdown/unlock

S. no.	*Lockdown/unlock*	*Average growth rate (Active cases)*	*Median growth rate (Active cases)*
1	Lockdown 1.0	1.06115606	1.04261554
2	Lockdown 2.0	1.00494153	1.00514210
3	Lockdown 3.0	0.98696387	0.98816720
4	Lockdown 4.0	1.02769516	1.03035692
5	Unlock 1.0	1.02928239	1.02869489

As we can see in Table 1, the active cases growth rate decreases in all lockdowns. In lockdown 3.0 the growth rate of active cases is very less as compared to another lockdown. The lockdowns are shown a slight effect on the growth rate of COVID-19

active cases in India for controlling the spread of this virus (Kalipe et al., 2018; Gupta and Pal, 2019).

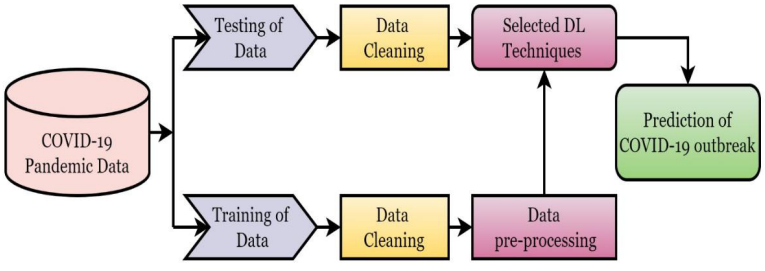
3.3 Data pre-processing

Data pre-processing method is used to remove unwanted and noisy data from the dataset. This method helps in arranging datasets in a proper format. The data structure is prepared for proper data analysis. Data cleaning is required that plays the main role in data pre-processing before modelling (Liu et al., 2021). The data contain various missing values this cause an error when this file is used directly as an input file. Therefore, the missing values fill as NA in the input file. The data pre-processing method is used on the dataset and then the ML technique is applied to it (Josephine et al., 2021; Kelly et al., 2019). Once the dataset is in preparing a proper format then it needs to save .CSV or .XLSX file format. Now we can perform various ML techniques to prepare a dataset for prediction (Khan et al., 2021; MacIntyre et al., 2021).

4 Methodology

The predictive model provides various solutions for a pandemic by using ML techniques (Kalipe et al., 2018; Hanna and Radwan, 2020). With the help of these techniques, we perform parameterised optimisation to decrease the RMSE score of the model. The proposed outbreak prediction model has four modules as illustrated in Figure 6 and they are described below:

Figure 6 A framework for COVID-19 outbreak prediction model (see online version for colours)



This section describes the methodology and details of the work and models for the prediction of COVID-19 cases. The main objective of this research is to forecast the number of confirmed cases, deaths and cured cases for the next coming week. The task helps to gather knowledgeable information from COVID-19 data that helps in tackling the pandemic. Various models were applied to get the maximum results to perform the task. Based on the results, the models provide various solutions for other countries to understand the spread rate of the virus and take the necessary steps to overcome this pandemic. It also helps countries to prepare for the next coming situation of the pandemic. The main works done in this research are as follows-mainly these tasks were performed they are as follows:

- The COVID-19 India cases forecasting for the next coming week using different models for the number of confirmed cases, deaths and recoveries cases.
- Prediction of COVID-19 India confirmed deaths and cured cases using the best-performed model. Comparing the performance of models based on RMSE score among all the models.

In this research, we use the logistic curve sigmoid model, FB prophet model, SARIMAX model, SVM model, LR model and PR model to accomplish the above task.

4.1 Sigmoid model

The sigmoid model is also known as the logistic function. It works as an activation function to add nonlinearity to the ML model. The S-shaped curve model is generated in a logistic curve. The main purpose of the sigmoid function is to identify a number of parameter values to pass as outputs to the ML model. To select values, the function must be nonlinear in those parameters (Wannier et al., 2019).

For fitting data into sigmoid function $S(p)$, the following equation (3) is used.

$$S(p) = \frac{M}{1 + e^{-g(p-p_0)}} \tag{3}$$

where

p and p_0 is the sigmoid midpoint

M is the curve maximum value

g is the logistic growth rate.

The following equation is used to select the values to measure the respective effect of the parameter.

4.2 Facebook prophet model

Facebook prophet model is an open tool provided by Facebook for time series forecasting. Forecasting helps to predict the cases and understand its trend. The model is based on the catalytic additive model with the most suitable nonlinear propensity for weathering and is also used to measure the effect of time (Hojeong and Songhee, 2020). The forecasting is based on current and historical time series data. To measure the tendency there are four components of time series:

- trend: trend helps in data to increase and decrease its trend over a period and helps to filter out seasonal variation.
- Seasonality: it is a short period time of variation that occurs during the seasonality.

The Facebook prophet model is used for the prediction of the current trend of seasonal and timestamp analysis. The equation used for the model is as follows in equation (4).

$$F(t) = p(t) + q(t) + r(t) + h(t) \tag{4}$$

where

$p(t)$ is trend that vary over long period of time

$q(t)$ is seasonality that refer short term changes

$r(t)$ is the time stamp for forecasting

$h(t)$ is unmapped changes that depend upon the any circumstances

$F(t)$ is the forecasting of cases.

4.3 SRIMAX model

Seasonal autoregressive integrated moving average exogenous (SRIMAX). This is an extended version of the ARIMA model. ARIMA model is the combination of auto regression (AR) and moving average (MA). But the problem with the ARIMA model is that it cannot handle seasonality. To overcome this problem the extended version of ARIMA is designed is SRIMAX. The SRIMAX model is a combination of seasonality, regression and MAs. It is one of the most effective learning algorithms used for time series forecasting (Goldstein et al., 2011).

The SARIMA is a predecessor of the SARIMAX model. Shorthand notation of the SARIMA model is as follows in equation (5).

$$\text{SARIMA}(p, q, r)\times(P, Q, R, S) \tag{5}$$

where p is non-seasonal autoregressive, (AR) order, q is a non-seasonal difference, r is non-seasonal MA order, P is seasonal AR order, Q is a seasonal difference, R is seasonal MA order, and S is the length of the repeating seasonal pattern. Adding the AR and MA components solves the seasonality problem.

The SARIMAX model uses the exiting SARIMA model by adding the exogenous variable framework. To measure the record, the SARIMAX model of the time variation (C_t) in equation (6) are (1, 0, 0) (2, 0, 0, 10) looks like.

(6)

where C_{t-1}, C_{t-10}, C_{t-11}, $C_{t-20} \wedge C_{t-21}$ are seasons, ϵ_t is an exogenous variable, Φ_1, Φ_{10}, Φ_{11}, $\Phi_{20} \wedge \Phi_{21}$ are coefficient .

SRIMAX model is the most effective model to work with seasonality and exogenous variable.

4.4 Support vector machine (SVM) model

SVM is a ML approach for representing a training dataset (Wang et al., 2005). Each data point in the dataset is plotted in space. These points determine the number of characteristics that can be extracted from a set of coordinates (Chen, 2015). The points are divided into classes, each of which differs from the others in terms of the hyperplane's maximum distance from the point. The following are the SVM equations (7) and (8).

$$Min: \frac{1}{2} \| \omega \|_2^2 + C \sum_{i=1}^{m} \xi_i \tag{7}$$

$$\begin{aligned} &\text{Subjected}: \gamma_i\left(\omega \times x_i\right)+b \geq 1-\xi_i, \\ &\xi_i \geq 0, i=1, ..., m, \end{aligned} \tag{8}$$

C stands for regularisation parameter, b is coefficient, ξ_i stand for penalising relaxation variables, (x_i, γ_i) are training data for i = 1, 2, …, m, where m is a number of observations.

4.5 LR model

LR is a type of ML technique that defines the relationship between input and output variables (Dhamodharavadhani et al., 2020). Changes in the input variable have an immediate effect on the output variable (Françoise et al., 2021; Nazeri et al., 2021). A sloped straight line represents the relationship between the variables. A simple linear equation (t) in equation (9).

$$t = a\_0 + a\_1 * r \tag{9}$$

where r is the input variable, t is the output for the set value, and a_0 and a_1 are the model coefficients.

4.6 PR model

PR is a type of LR in which the data is fitted with a polynomial equation that has a curved relationship between the target and independent variables. The value of the target variable fluctuates in a non-uniform manner concerning the predictor in a curvilinear connection (s). With a single predictor, we obtain the following equation (10) in LR:

$$Y = \theta_0 + \theta_1 x \tag{10}$$

In the regression equation, θ is weight, Y represents the target, x represents the predictor, 0 represents the bias, and 1 represents the weight.

A linear relationship can be represented using this linear equation. The polynomial equation of degree 'n' is shown in equation (11).

$$Y = \theta_0 + \theta_1 x + \theta_1 x^2 + \theta_3 x^3 + \theta_4 x^4 + ... + + \theta_n x^n \tag{11}$$

Here, 0 is the bias, 1, 2, …, n are the weights in the PR equation x^2, x^3, …, x^n predictor and n is the degree of the PR.

As the value of n increases, the number of higher-order terms increases, making the equation more difficult.

5 Experimental results

In this research, the result has been achieved by applying the above model. The detailed experimental result of the mentioned work is reported here. We have also calculated the RMSE score of each model. After calculating the RMSE score we compare the performance result of each model. In our proposed model we use the sigmoid function, Facebook prophet model and SRIMAX model for prediction. The dataset used for forecasting is the COVID-19 India dataset. In this data set, we have made predictions for COVID-19 cases. The prediction is based on the total number of confirmed cases, total

number of deaths and the total number of cured cases in India. The data set we used is the cases of COVID-19 from January 30, 2020, to August 2, 2021. The attributes of the data set are date, time, state/union territory, conformed, deaths and cured cases. The forecasting is done for the next 60 days for the total number of confirmed, deaths and cured cases.

The RMSE score is used to measure the performance of ML models. The formula for determining the RMSE score is shown in equation (12).

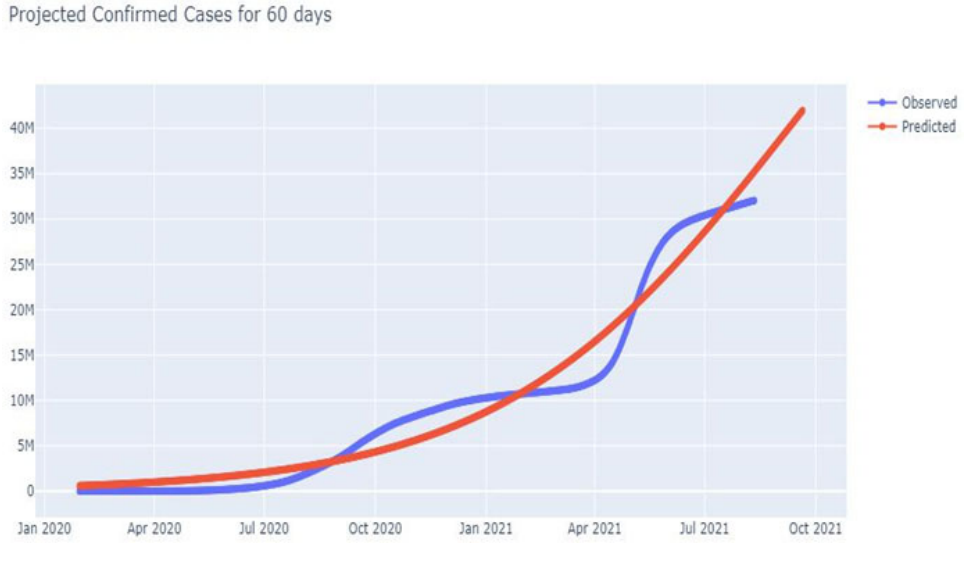
$$RMSE = \sqrt{N^1 \sum (Actual - Predicted)^2} \tag{12}$$

In *RMSE*, 'actual' and 'predicted' refer to predicted (output) and desired (goal) values, respectively. N is the number of data points used to evaluate the performance of ML models.

5.1 Forecasting using sigmoid model

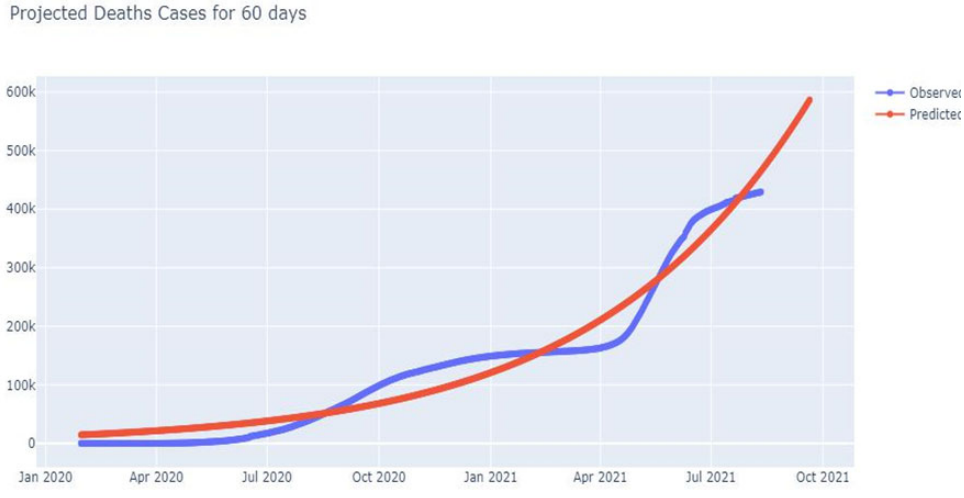
The cases of COVID-19 are increasing day by day in India, due to which the death and recovery rate is also very high. We are currently expecting a drop in cases and this situation can be best analysed using the sigmoid model. The experiment in Figure 8 describes the prediction of confirmed, death and cured cases of COVID-19 using the sigmoid model. In this experiment, we have predicted the confirmed, death and cured cases of COVID-19 for the next 60 days. We used CSV format data to fit the data in the sigmoid function. The data set used here is the COVID-19 India data set.

Figure 7 COVID-19 India, (a) confirmed, (b) deaths, (c) cured case prediction for next 60 days using sigmoid model (see online version for colours)



(a)

Figure 7 COVID-19 India, (a) confirmed, (b) deaths, (c) cured case prediction for next 60 days using sigmoid model (continued) (see online version for colours)



(b)

(c)

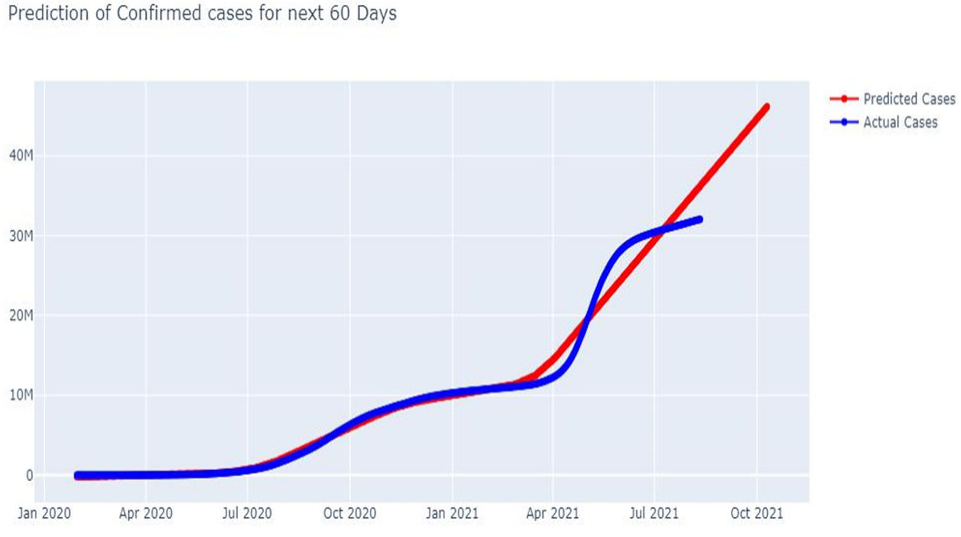
The graph shows that our model is fit to the data. We calculate the predicted cases. The increase in cases is predicted using the growth rate. We can see in the graph below that at some point in time the cases are falling but the predicted value is in a straight line. At some point, the graph shows a slight difference between the observed value and predicted value. The graph in Figure 7 describes the prediction of COVID-19 cases for the next 60 days.

5.2 Forecasting using Facebook prophet model

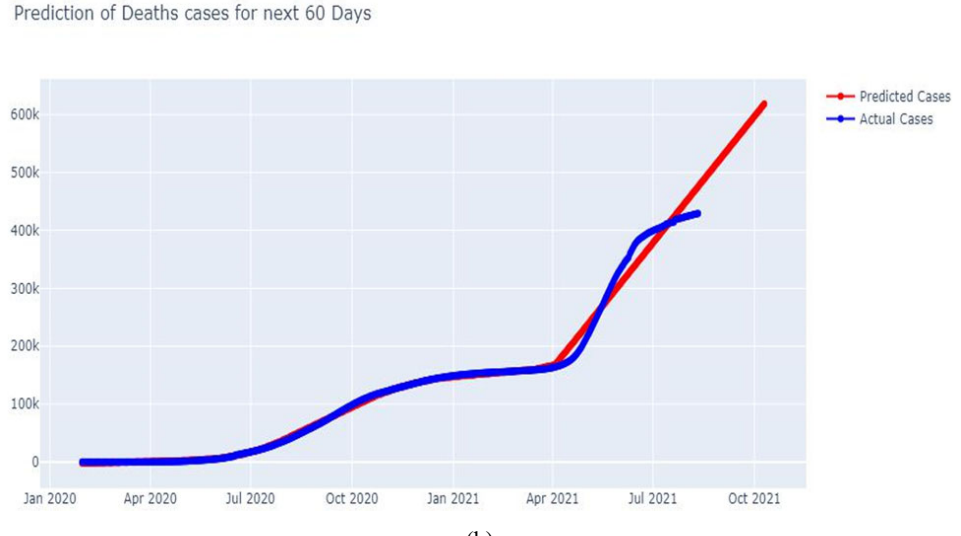
In the above experiment, we used the sigmoid function for forecasting but the estimation of the result is not as accurate as the real cases defined in the graph. The Facebook

prophet model is very well worked with seasonality data. As we all know the cases are very high in all seasons but at some time the cases fall. To measure this type of variation the Facebook prophet model is very well work with that variation. The graph in Figure 8 shows the prediction of confirmed, deaths and cured cases for the next 60 days.

Figure 8 COVID-19 India, (a) confirmed, (b) deaths, (c) cured case prediction for next 60 days using Facebook prophet model (see online version for colours)

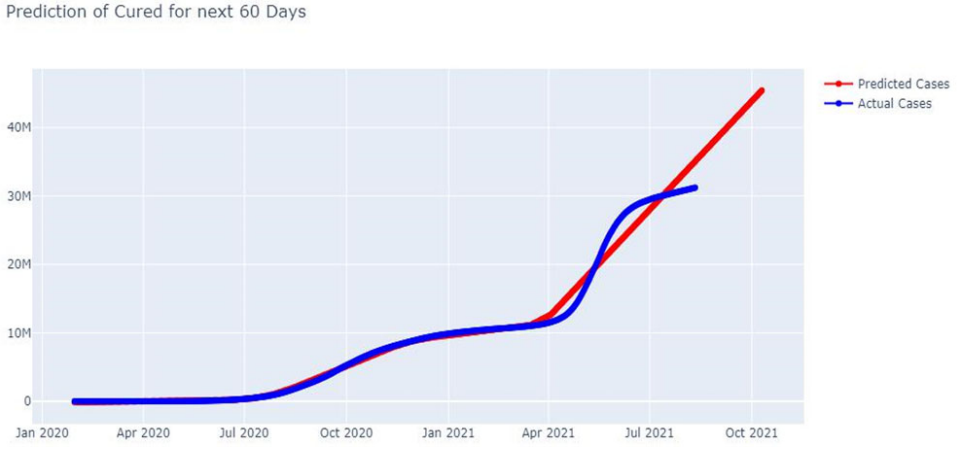


(a)



(b)

Figure 8 COVID-19 India, (a) confirmed, (b) deaths, (c) cured case prediction for next 60 days using Facebook prophet model (continued) (see online version for colours)

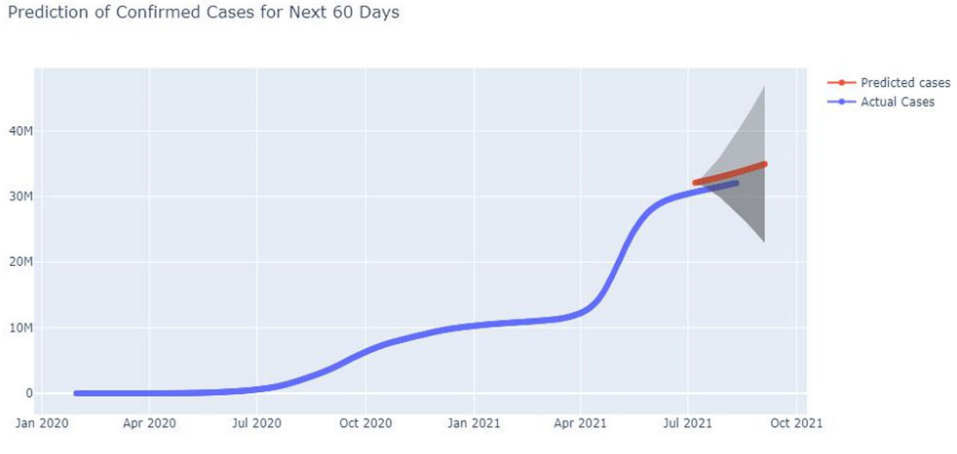


(c)

5.3 *Forecasting using SRIMAX model*

The SRIMAX model is an extension of the ARIMA model. As we know that COVID-19 is a global pandemic, the virus changes its form every time which we are not able to predict with simple ML models. SRIMAX model mainly performs for seasonal type data. In the experiment here we perform prediction of confirmed, deaths and cured cases of the COVID-19 pandemic with time-series data. In this model, we perform prediction for 60 days. Figure 9 shows the prediction graph using the SRIMAX model.

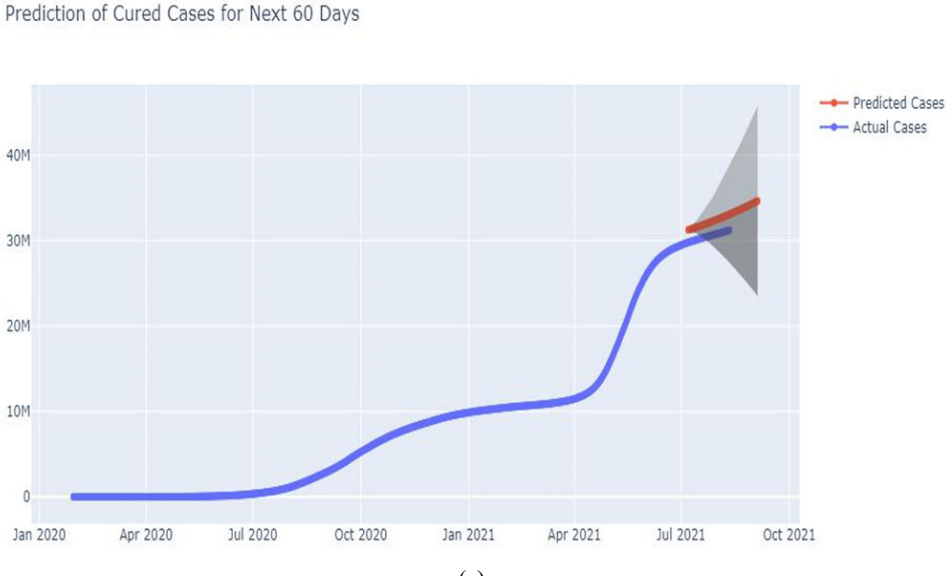
Figure 9 COVID-19 India, (a) confirmed, (b) deaths, (c) cured case prediction for next 60 days using SRIMAX model (see online version for colours)



(a)

Figure 9 COVID-19 India, (a) confirmed, (b) deaths, (c) cured case prediction for next 60 days using SRIMAX model (continued) (see online version for colours)

(b)



(c)

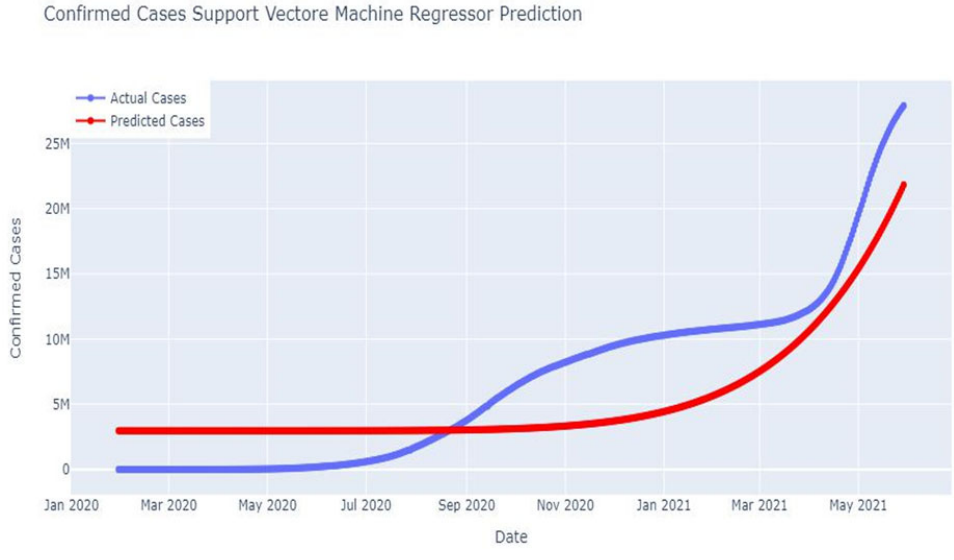
5.4 Forecasting using SVM model

With time-series data, the SVM model produces accurate results. As we all know, the number of instances is high throughout the year, but at some point, the number of cases decreases. The SVM model is particularly well suited to measuring this type of variance. In Figure 10, the prediction of confirmed, fatalities, and cured patients over the following 60 days is shown.

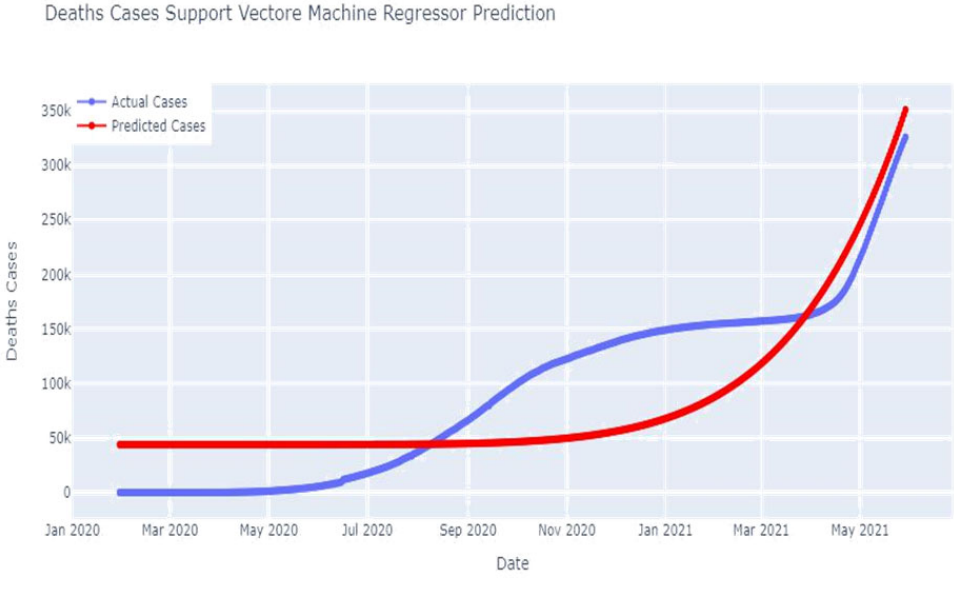
5.5 Forecasting using LR model

When employing time series data, the LR model produces accurate results. As we all know, the number of instances is high throughout the year, but it gradually decreases. The LR model is particularly well-suited to this type of variance. Over the next 60 days, Figure 11 shows the expected number of confirmed, fatal, and curable cases.

Figure 10 COVID-19 India, (a) confirmed, (b) deaths, (c) cured case prediction for next 60 days using SVM model (see online version for colours)

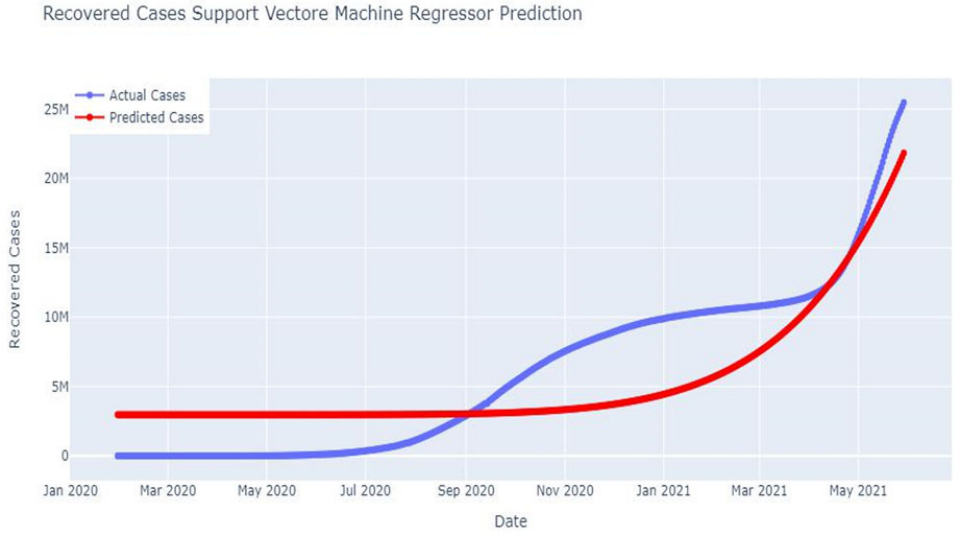


(a)



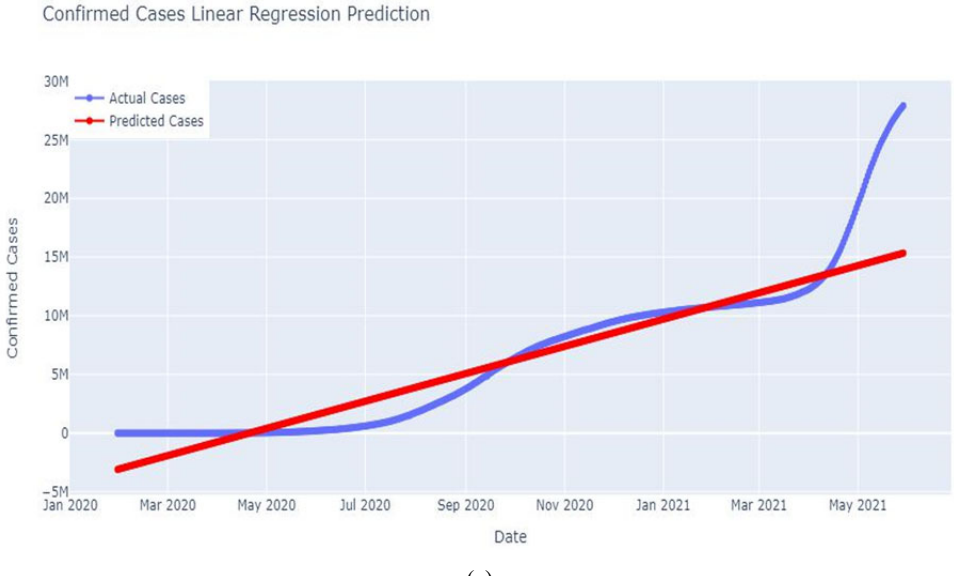
(b)

Figure 10 COVID-19 India, (a) confirmed, (b) deaths, (c) cured case prediction for next 60 days using SVM model (continued) (see online version for colours)



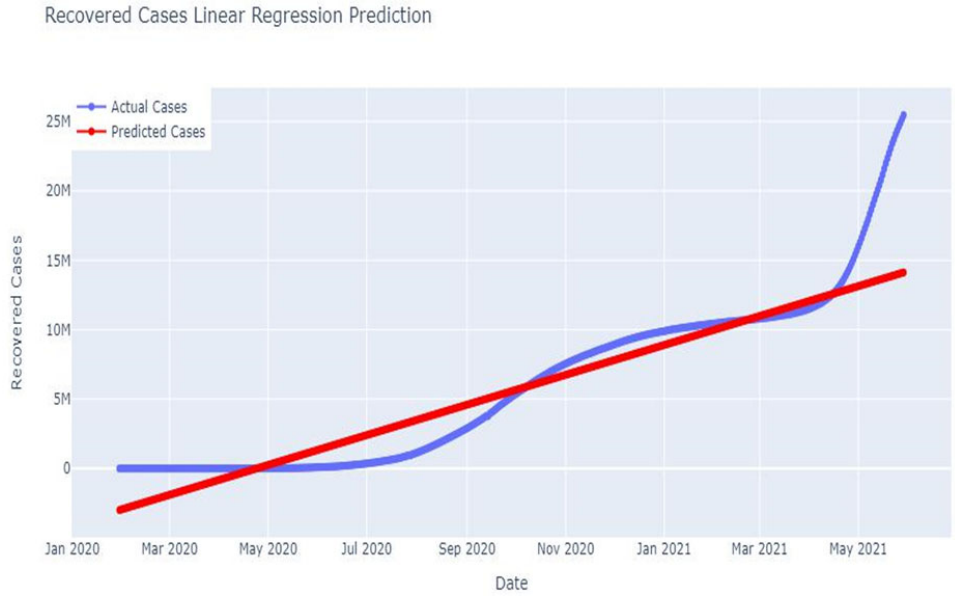
(c)

Figure 11 COVID-19 India, (a) confirmed, (b) deaths, (c) cured case prediction for next 60 days using LR model (see online version for colours)

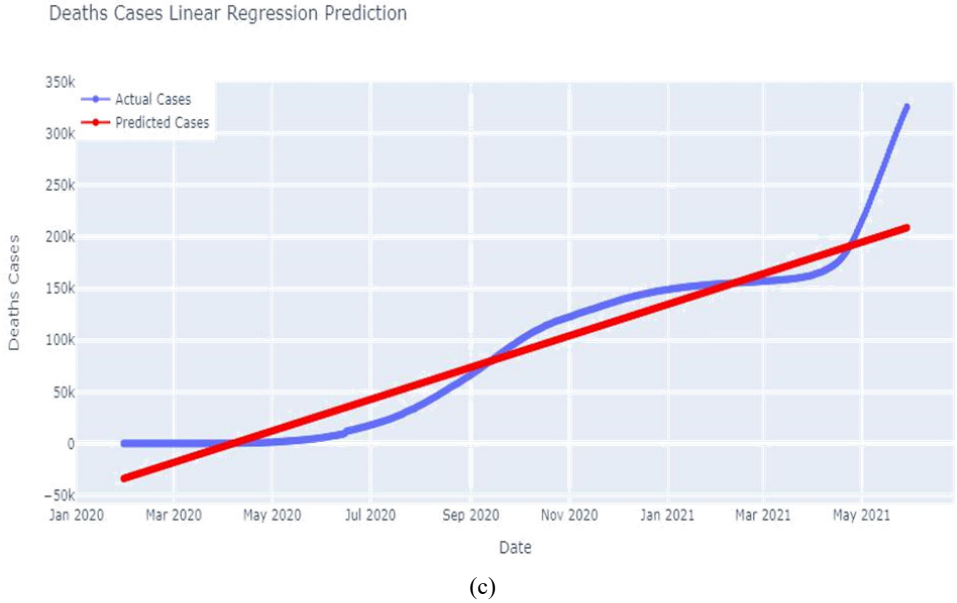


(a)

Figure 11 COVID-19 India, (a) confirmed, (b) deaths, (c) cured case prediction for next 60 days using LR model (continued) (see online version for colours)



(b)



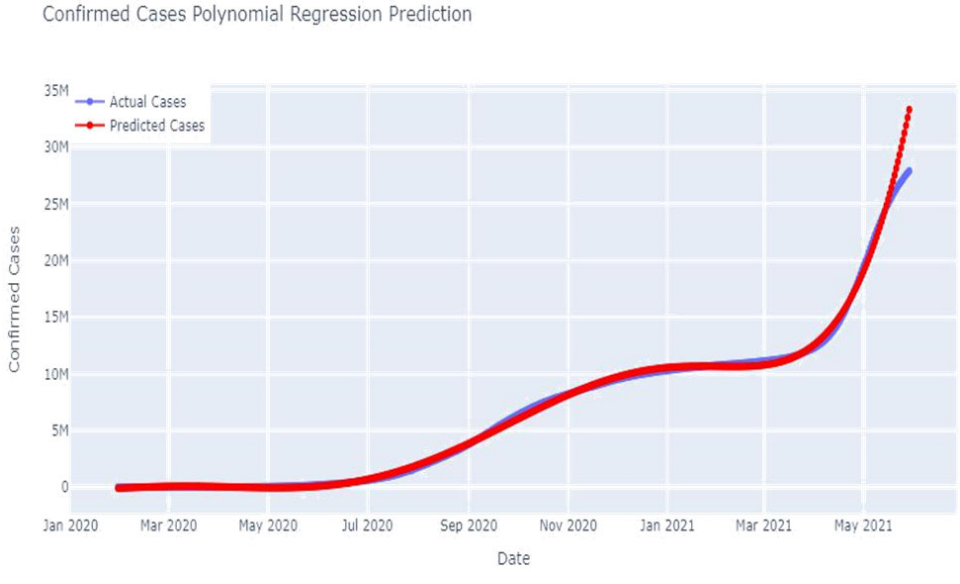
(c)

5.6 Forecasting using PR model

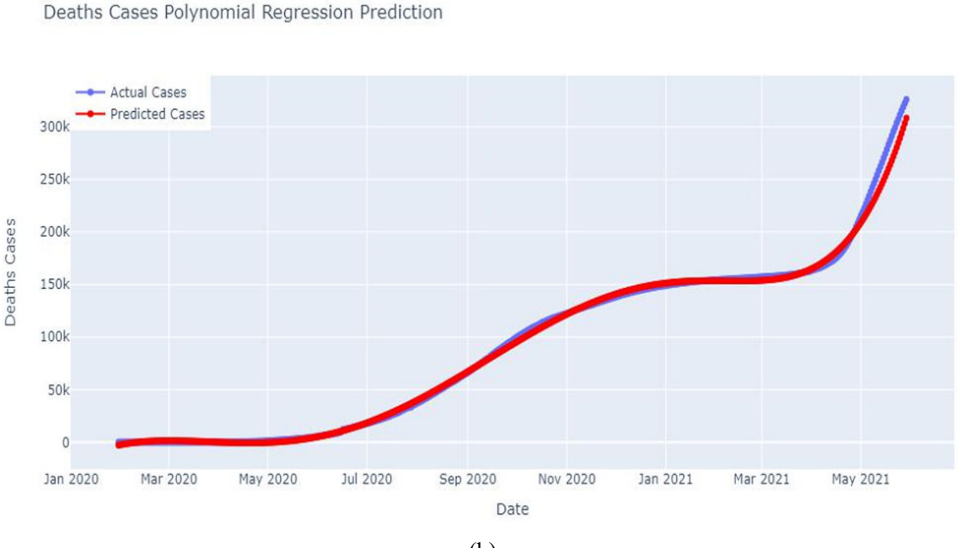
The PR model gives accurate results when using time series data. As we all know, the number of cases is large throughout the year, but it begins to decline at some time. This

form of volatility is particularly well-suited to the PR model. Figure 12 depicts the predicted number of confirmed, fatal, and cured cases over the next 60 days.

Figure 12 COVID-19 India, (a) confirmed, (b) deaths, (c) cured case prediction for next 60 days using PR model (see online version for colours)

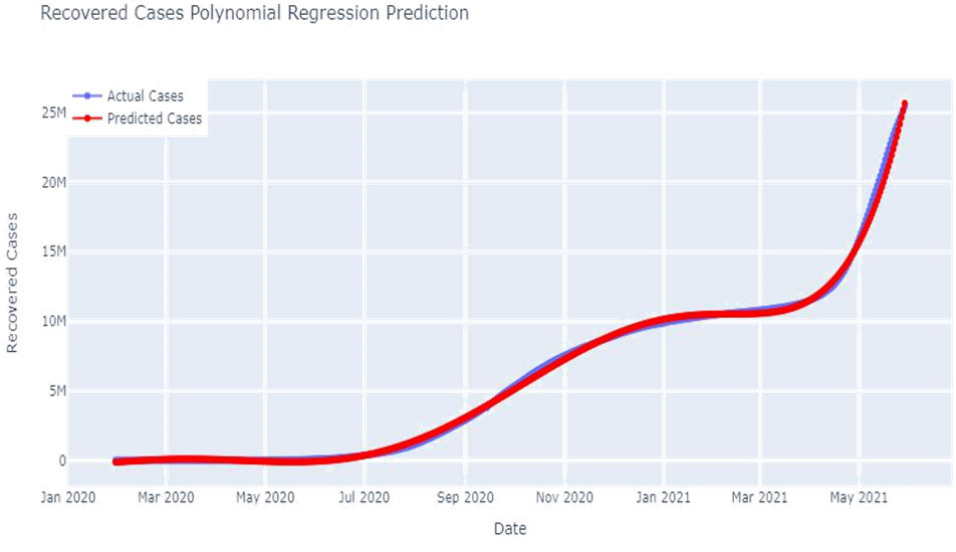


(a)



(b)

Figure 12 COVID-19 India, (a) confirmed, (b) deaths, (c) cured case prediction for next 60 days using PR model (continued) (see online version for colours)



(c)

5.7 Comparative analysis

As we can see in the above approach for predicting the number of confirmed, deaths and recovery cases in India among all these approaches, the visual representation of the prediction from Figure 12 is more accurate. After seeing the graph forecasting we can see that the PR graph shows more accurate results. Also, the RMSE score as highlighted in Table 2 confirms that the PR approach has a very less RMSE score as compared to other ML models. So, the PR model is the best to follow the growing trend.

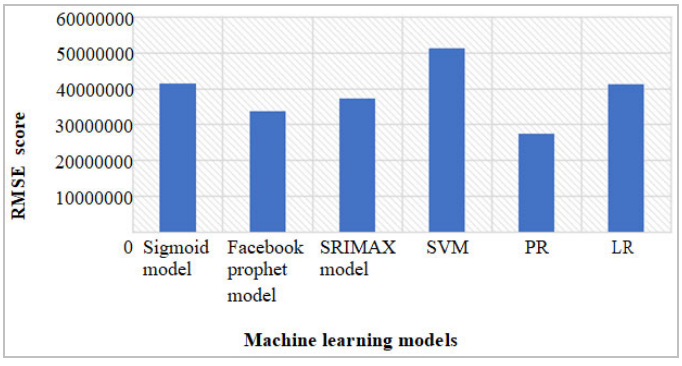
Table 2 Comparative results

Model	*RMSE score*
Sigmoid model	41,367,951.221521486
Facebook prophet model	33,629,589.416571088
SRIMAX model	37,118,234.28333981
SVM	51,237,564.239517636
PR	27,435,923.21693116
LR	41,141,511.296706144

The comparison analysis shows that the sigmoid model has the highest RMSE score in the comparison among all other models. Also, the graph of the sigmoid model shows less accurate prediction results when compared with other ML models.

As we can see in Figure 13, PR model has a very less RMSE score. It means that the prediction result is more accurate in comparison to other models. So, for future prediction of a pandemic like COVID-19 we can refer to that model. Also, the forecasting trend line of the PR prediction model is more accurate compared with other models.

Figure 13 Comparison of RMSE score of models (see online version for colours)



6 Conclusions and future work

The SARS-CoV2 (COVID-19) virus has taken over the world. Early detection of the transmission can aid in the implementation of appropriate measures. Employing data from the Johns Hopkins dashboard, this research recommended using ML and deep learning models for epidemic analysis. In anticipating the COVID-19 transmission, the results demonstrate that PR produced the lowest RMSE score when compared to other ML models. It indicates that the forecast result is more accurate when compared to other models. However, if the spread follows the PR model's expected trend, it will result in a massive loss of life due to the exponential expansion of transmission over the world. As a result, we may use that model to forecast future pandemics such as COVID-19. COVID-19 growth can be slowed and stopped by limiting the number of susceptible individuals among infected persons, as seen in India.

ML forecasting models will be utilised in the future to improve prediction results for COVID-19 situations. In addition, for modelling the highly accurate prediction model for the ongoing COVID-19 pandemic, this study will include numerous elements or variables such as population, immunity, climate like temperature, humidity, wind speed, rainfall, and vaccine factors. Further, using ML we can provide various solutions to tackle this type of pandemic.

References

Abirami, S. and Chitra, P. (2020) 'Energy-efficient edge based real-time healthcare support system', in *Advances in Computers*, Vol. 117, No. 1, pp.339–368, Elsevier.

Ardabili, S.F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A.R., Reuter, U. and Atkinson, P.M. (2020) 'Covid-19 outbreak prediction with machine learning', *Algorithms*, Vol. 13, No. 10, pp.244–249.

Asher, J. (2018) 'Forecasting Ebola with a regression transmission model', *Epidemics*, Vol. 22, No. 1, pp.50–55.

Chen, W. (2015) *Modelling the Logisitcs Response to Disasters*, Doctoral dissertation, Lyon, INSA, Vol. 4, No. 8, pp.112–119.

Dhamodharavadhani, S., Rathipriya, R. and Chatterjee, J.M. (2020) 'COVID-19 mortality rate prediction for India using statistical neural network models', *Frontiers in Public Health*, Vol. 8, No. 28, p.441.

Françoise, K., Daniele, P., Atte, A., Laurent, M., Aymeric, F.D.H., Andreas, H., Christophe, L., Jorge, G., Alexander, S. and Stefano, M. (2021) 'Modelling COVID-19 dynamics and potential for herd immunity by vaccination in Austria, Luxembourg and Sweden', *Journal of Theoretical Biology*, December, Vol. 530, No. 1, p.110874.

Goldstein, E., Cobey, S., Takahashi, S., Miller, J.C. and Lipsitch, M. (2011) 'Predicting the epidemic sizes of influenza A/H1N1, A/H3N2, and B: a statistical method' *PLoS Med*., Vol. 8, No. 7, pp.95–101.

Gupta, R. and Pal, S.K. (2020) 'Trend analysis and forecasting of COVID-19 outbreak in India', *MedRxiv*.

Hanna, W.K. and Radwan, N.M. (2020) 'Heart disease patient risk classification based on neutrosophic sets', *International Journal of Business Intelligence and Data Mining*, Vol. 20, No. 1, pp.93–106.

Hojeong, P. and Songhee, H.K. (2020) 'A study on herd immunity of COVID-19 in South Korea: using a stochastic economic-epidemiological model', *Environmental and Resource Economics*, July, Vol. 76, pp.665–670.

Iwendi, C., Bashir, A.K., Peshkar, A., Sujatha, R., Chatterjee, J.M., Pasupuleti, S., Mishra, R., Pillai, S. and Jo, O. (2020) 'COVID-19 patient health prediction using boosted random forest algorithm', *Frontiers in Public Health*, Vol. 8, No. 1, p.357.

Josephine, N.A.T., Van, K.N. and Hernandez-Vargas, E.A. (2021) 'Network models to evaluate vaccine strategies towards herd immunity in COVID-19', *Journal of Theoretical Biology*, December, Vol. 531, No. 1, p.110894.

Kalipe, G., Gautham, V. and Behera, R.K. (2018) 'Predicting malarial outbreak using machine learning and deep learning approach: a review and analysis', in *2018 International Conference on Information Technology (ICIT)*, IEEE, pp.33–38.

Kelly, J.D., Park, J., Harrigan, R.J., Hoff, N.A., Lee, S.D., Wannier, R. and Schoenberg, F.P. (2019) 'Real-time predictions of the 2018–2019 Ebola virus disease outbreak in the Democratic Republic of the Congo using Hawkes point process models', *Epidemics*, Vol. 28, No. 1, pp.782–789.

Khan, M.A., Abidi, W.U.H., Al Ghamdi, M.A., Almotiri, S.H., Saqib, S., Alyas, T. and Mahmood, N. (2021) 'Forecast the influenza pandemic using machine learning', *CMC-Computers Materials & Continua*, Vol. 66, No. 1, pp.331–357.

Liu, X.X., Hu, S., Fong, S.J., Crespo, R.G. and Herrera-Viedma, E. (2021) 'Modelling dynamics of Coronavirus disease 2019 spread for pandemic forecasting based on Simulink', *Physical Biology*, Vol. 18, No. 4, p.045003.

MacIntyre, C.R., Costantino, V. and Trent, M. (2021) 'Modelling of COVID-19 vaccination strategies and herd immunity, in scenarios of limited and full vaccine supply in NSW, Australia', *Vaccine*, April.

Miranda, G.H., Baetens, J.M., Bossuyt, N., Bruno, O.M. and De Baets, B. (2019) 'Real-time prediction of influenza outbreaks in Belgium', *Epidemics*, Vol. 28, No, 1, pp.78–85.

Muhilthini, P., Meenakshi, B.S., Lekha, S.L. and Santhanalakshmi, S.T. (2018) 'Dengue possibility forecasting model using machine learning algorithms', *Int. Res. J. Eng. Technol.*, Vol. 5, No. 3, pp.1661–1665.

Nazeri, A., Ghareh Gozlu, H., Faraji, F. and Asakareh, S. (2021) 'Analysis of road accident data and determining affecting factors by using regression models and decision trees', *International Journal of Business Intelligence and Data Mining*, Vol. 18, No. 4, pp.449–471.

Nazir, A. and Khan, R.A. (2021) 'A novel combinatorial optimization-based feature selection method for network intrusion detection', *Computers & Security*, Vol. 102, No. 1, p.102164.

Punn, N.S., Sonbhadra, S.K. and Agarwal, S. (2020) 'COVID-19 epidemic analysis using machine learning and deep learning algorithms', *MedRxiv*, Vol. 1, No. 1, pp.143–149.

Rongali, S. and Yalavarthi, R. (2017) 'An improved ant colony optimization for parameter optimization using support vector machine', *Int. J. Eng. Adv. Technol. (IJEAT)*, Vol. 6, No. 3, pp.198–204.

Tian, Y., Luthra, I. and Zhang, X. (2020) 'Forecasting COVID-19 cases using machine learning models', *Medrxiv*, Vol. 1, No. 1, pp.121–125.

Wang, J., Wu, X. and Zhang, C. (2005) 'Support vector machines based on K-means clustering for real-time business intelligence systems', *International Journal of Business Intelligence and Data Mining*, Vol. 1, No. 1, pp.54–64.

Wannier, S.R., Worden, L., Hoff, N.A., Amezcua, E., Selo, B., Sinai, C. and Kelly, J.D. (2019) 'Estimating the impact of violent events on transmission in Ebola virus disease outbreak, Democratic Republic of the Congo', *Epidemics*, Vol. 28, No. 1, pp.1132–1141.

Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G. and Zhang, Y.Z. (2020) 'A new coronavirus associated with human respiratory disease in China', *Nature*, Vol. 579, No. 7, pp.265–269.

Xu, X., Jiang, X., Ma, C., Du, P., Li, X., Lv, S. and Li, L. (2020) 'A deep learning system to screen novel coronavirus disease 2019 pneumonia', *Engineering*, Vol. 6, No. 10, pp.1122–1129.

Yadav, M., Perumal, M. and Srinivas, M. (2020) 'Analysis on novel coronavirus (COVID-19) using machine learning methods', *Chaos, Solitons & Fractals*, Vol. 139, No. 1, pp.561–566.

Websites

https://covid19.who.int/info/.