Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Predicting the impact of the third wave of COVID-19 in India using hybrid statistical machine learning models: A time series forecasting and sentiment analysis approach

Sumit Mohan[a,*], Anil Kumar Solanki[a], Harish Kumar Taluja[b], Anuradha[c], Anuj Singh[d]

[a] *Department of Computer Science and Engineering, Bundelkhand Institute of Engineering and Technology, Jhansi, AKTU, Lucknow, India*
[b] *Department of Computer Science and Engineering, Noida International University, Noida, India*
[c] *Department of Computer Science and Engineering, Ajay Kumar Garg Engineering College, Ghaziabad, AKTU, Lucknow, India*
[d] *Department of Computer Science and Engineering, Kamla Nehru Institute of Technology, Sultanpur, AKTU, Lucknow, India*

ARTICLE INFO

Keywords:
ARIMA
COVID-19
Prophet
Natural language processing
Sentiment analysis
Time series forecasting

ABSTRACT

Background: Since January 2020, India has faced two waves of COVID-19; preparation for the upcoming waves is the primary challenge for public health sectors and governments. Therefore, it is important to forecast future cumulative confirmed cases to plan and implement control measures effectively.

Methods: This study proposed a hybrid autoregressive integrated moving average (ARIMA) and Prophet model to predict daily confirmed and cumulative confirmed cases. The built-in auto.arima function was first used to select the optimal hyperparameter values of the ARIMA model. Then, the modified ARIMA model was used to find the best fit between the test and forecast data to find the best model parameter combinations. Articles, blog posts, and news stories from virologists, scientists, and health experts related to the third wave of COVID-19 were gathered using the Python web scraping package Beautiful Soup. Their opinions (sentiments) toward the potential third wave were analyzed using natural language processing (NLP) libraries.

Results: A spike in daily confirmed and cumulative confirmed cases was predicted in India in the next 180 days based on past time series data. The results were validated using various analytical tools and evaluation metrics, producing a root mean square error (RMSE) of 0.14 and a mean absolute percentage error (MAPE) of 0.06. The NLP processing results revealed negative sentiments in most articles and blogs, with few exceptions.

Conclusion: The findings of this study suggest that there will be more active cases in the upcoming days. The proposed models can forecast future daily confirmed and cumulative confirmed cases. This study will help the country and states plan appropriate public health measures for the upcoming waves of COVID-19.

1. Introduction

The coronavirus disease 2019 (COVID-19) results from infection with the SARS-CoV-2 virus. The first case of COVID-19 was reported in Wuhan, China, in December 2019, and the global fight against the virus continues nearly two years later. According to India's daily health bulletin, over 33,000 cases were reported in the country on September 11, 2021. Compared to the previous week's data (September 4, 2021), this represents a slight drop in new cases. The Ministry of Health data shows that the number of active COVID-19 cases has dropped to its lowest level in the last four months, indicating that the second wave is ending. The rates of deaths and active cases have fallen since July 2021, and the country has had the lowest number of active cases since March 2020. According to a Reuters poll of medical experts, the third wave of COVID-19 will strike India by the end of 2021 [1]. The pandemic will continue to be a public health concern for at least another year. A substantial increase in vaccines will likely provide some protection against a new pandemic, according to a poll of 40 healthcare professionals and scientists from around the world, over 85% of whom expected the third wave would come in late November or December. However, more than 70% of experts agreed that any future COVID-19 pandemic would have less impact than the current one. The second wave of COVID-19 resulted in many deaths due to shortages of vaccines, treatments, oxygen, and hospital beds. Dr. Randeep Guleria, head of the All India Institute of

* Corresponding author.
E-mail addresses: sumitmohan@bietjhs.ac.in (S. Mohan), aksolanki@bietjhs.ac.in (A.K. Solanki), harishtaluja@gmail.com (H.K. Taluja), anutaluja@gmail.com (Anuradha), anuj.2295@knit.ac.in (A. Singh).

https://doi.org/10.1016/j.compbiomed.2022.105354
Received 9 October 2021; Received in revised form 24 February 2022; Accepted 24 February 2022
Available online 26 February 2022

Table 1
Vaccination progress as of February 13, 2022.

Groups	1st Dose	2nd Dose	Precaution Dose
Healthcare Workers	1,03,99,410	99,30,634	38,78,308
Frontline Workers	1,84,05,152	1,73,74,818	53,58,037
Age Group 15–18 Years	5,20,32,858	1,47,92,245	N/A
Age Group 19–44 Years	54,80,44,294	42,63,39,386	N/A
Age Group 45–59 Years	20,16,19,377	17,62,74,802	N/A
Age Group ≥ 60 Years	12,58,81,409	10,98,24,107	79,94,610

Table 2
Summary of recent related works.

Approach/Model	Country	Accuracy	Reference
Genetic programming/ gene expression programming	Australia	Genetic programming better than other ML models	Salgotra et al., 2021 [35]
Deep learning/ ARIMA, LSTM, and SLSTM	India/Chennai	SLSTM better than LSTM and ARIMA	Devaraj et al., 2021 [36]
ARIMA, KNN, R.F., SVM, Holt-Winters, SARIMA, PR, decision trees	Bulgaria, Greece, Russia, China, Iran, Sweden, India, The Netherlands	Holt-Winters, SARIMA better than other ML models	Saba et al., 2021 [37]
SARIMA, LSTM, ARIMA, and RF	Spain, India, USA, Worldwide	SARIMA and LSTM Better than ARIMA and RF	Malki et al., 2021 [38]
LR, SARIMAX, SSL, statistical SARIMAX	India, China, Brazil, USA	SSL better than others	Patil et al., 2021 [39]
Uncertain time series forecasting	China	Better than traditional time series forecasting	Ye et al., 2021 [29]
VARIMAX	Philippines	Able to forecast future cases with ordinary least squares algorithm	Jamdade et al., 2021 [31]
ARIMA and SARIMA	Top 16 infected countries: Brazil, Chile, India, Colombia, Russia, Mexico, Iran, Peru, Bangladesh	SARIMA models outperform the ARIMA models	Arun et al., 2021 [30]
ARIMA, Holt-Winters, TBATS, and Spline	USA, Italy	ARIMA and Holt-Winters better than TBATS and Spline	Gecili et al., 2021 [32]
Epidemiology SIR with regression, ARIMA, and Prophet	Top 20 countries	SEIR better for long term prediction, and POLY d(2) better for short periods	Furtado et al., 2021 [9]
ARIMA	Bangladesh	ARIMA (0,2,1) and ARIMA (0,1,1) better than others	Kundu et al., 2021 [12]
ARIMA	Egypt	ARIMA (2,1,2) and ARIMA (2,1,3)	Sabry et al., 2021 [16]
ARIMA	India	ARIMA (2,2,2)	Roy et al., 2021 [33]

Medical Science, has stated that upcoming waves would be better controlled and result in fewer cases. Furthermore, India will achieve herd immunity at the end of this year because of the government's mass vaccination campaign, as vaccination provides some protection against COVID-19 [2].

According to official data from the healthcare ministry (Table 1), approximately 96 crore people received at least one vaccination as of February 13, 2022. In contrast, India has reached the historic milestone of administering 150 crore vaccine doses on January 7, 2022. Healthcare experts and scientists have stated that children under the age of 18 would be the most susceptible during the third wave because they are the least vaccinated population. Now the government has started a vaccination campaign for them. Robert Gallo has posited that India would surely gain herd immunity through vaccination and virus exposure by 2023. However, a new coronavirus variant will challenge the global health system and World Health Organization (WHO). Therefore, upcoming waves are the main concern for the government and public health sector. To address this concern, this study predicts the impact of the third wave in India by examining future cases using time series forecasting techniques. We discovered patterns within and relationships between the active and recovered cases and related deaths by analyzing COVID-19 data from January 2020 to August 2021 (Fig. 1). Most cases in the first and second waves occurred between December and January and May/June. Figs. 1 and 2 show a link between active and cured cases and deaths. These correlations and patterns help to build an understanding of the time series data for analyzing COVID-19. We used the ARIMA and Prophet models to analyze COVID-19 time series data to reveal these correlations and patterns. In this study, data from January 30, 2020 to August 11, 2021 was collected from the website of Covid19india (see Table 2).

2. Materials and methods

The COVID-19 pandemic is the primary global concern because every country is fighting the coronavirus, with scientists and health experts continuously working to overcome its impacts. Since December 2019, every country has faced ongoing waves of COVID-19; thus, predicting future cases based on past data using various models or algorithms has become a focus of recent research. Making accurate and reliable forecasts is the main challenge of this task because the number of COVID-19 cases varies between countries. Seasonality and trends are the essential parameters of time series forecasting models for predicting future cases of COVID-19. Susceptible-Infectious-Recovered model (SIR)-type models are often used to visualize curves for diseases, like COVID-19, in an epidemic.

Time series–based models are widely used to forecast future cases based on past cases. Gaur (2020) analyzed the data of almost 20 countries and compared the forecast results of several models, namely the Susceptible-Exposed-Infectious-Recovered (SEIR), autoregressive integrated moving average (ARIMA), Prophet, and polynomial regression models. The results generated by ARIMA and SEIR were found to be reliable for long-term predictions, whereas polynomial regression was more appropriate for short-term predictions (up to 3 weeks) [3]. The authors further conducted automated fitting, parameter optimization, and what-if analysis using the SEIR model for current and future data, as new cases were increasing exponentially in Indonesia. According to the researchers, their findings will support planning by public healthcare bodies. A study conducted in Iran used the ARIMA model to predict the daily total active cases and found an increasing trend in confirmed cases. Data from February 20 to May 04, 2020, were used to predict future cases in Iran [4]. Another study used three deep learning models to forecast COVID-19 cases. Abbasimehr (2021) proposed such models for short and long-term forecasting [5]. Another study was conducted on data from March 16 to May 17, 2020, in India. ARIMA (2,3,1), (2,2,0), and (1,3,1) were found better for the long-term forecasting [6]. Perone (2020) conducted a study in Italy to predict cases after April 4, 2020, using data from February 20 to April 4, 2020 [7]. Ghosal (2020) used a linear regression machine learning model to forecast future COVID-19 deaths in India [8]. Another study was conducted in China used the epidemiology-SIR with regression, ARIMA, and Prophet. Furtado (2021) found that SEIR performed better for long-term forecasting and polynomial regression of degree 2 was better for short-term forecasting [9]. Fanoodi (2019) utilized ARIMA and exponential smoothing to predict the demands of blood platelets using data from 2013 to 2018 [10]. ARIMA (0,2,1), (1,2,0), and (1,2,1) were found to be the models with the most effective parameters. In the case of time series forecasting, the lowest root mean square error (RMSE) value was preferable [11]. When the RMSEs of the two models were combined, the model performed better.

Kundu (2021) used data from March 8 to October 16, 2020, to

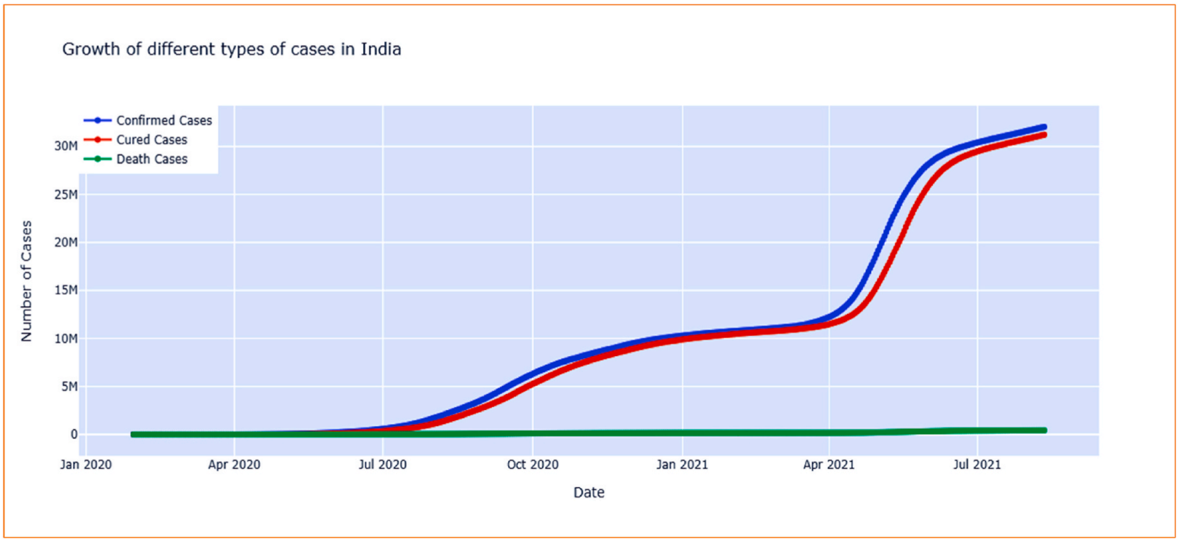

Fig. 1. Comparison of active and recovered cases and deaths.

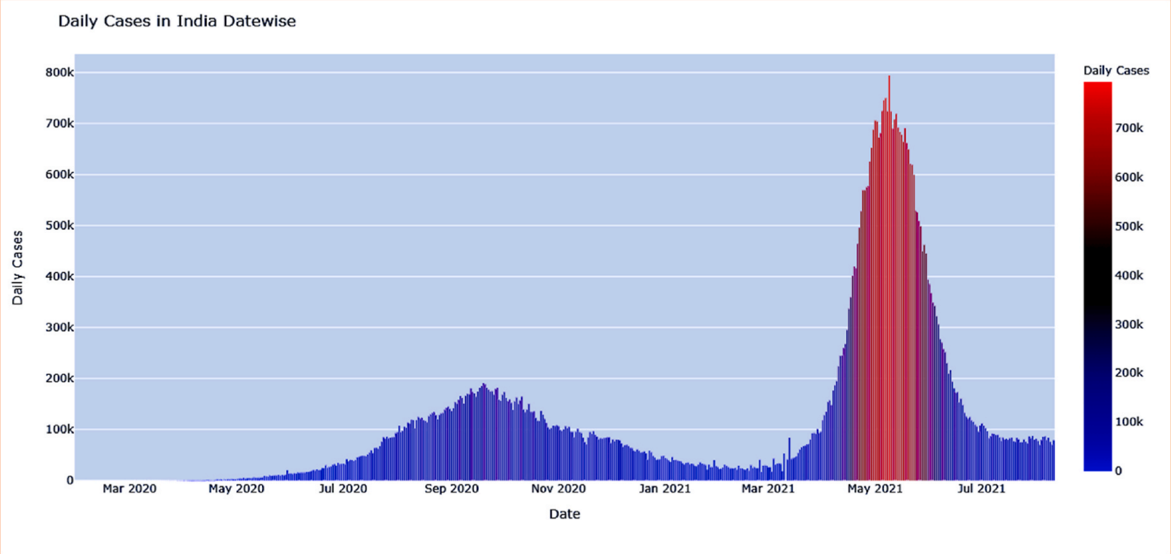

Fig. 2. Daily confirmed cases in India, March 2020–August 2021.

predict future cases from October 17 to November 15, 2020, in Bangladesh. ARIMA (2,3,1) and (1,3,1) were used to predict future cases. The results showed that there would be fewer or the same number of cases in the following month [12]. Building an effective forecasting model for predicting the future development of an infectious disease requires time series analysis; such a model might play an important role in predicting future cases. Parbat (2020) used a support vector regression model to predict future cases in India using data from March to April 2020. The accuracy of the proposed model was 97% for deaths and 87% for daily new cases [13]. Bayyurt (2020) proposed a study using time series forecasting and machine learning models to forecast future cases [14]. The ARIMA and Holt-Winters time series forecasting models were able to predict the next 20 days of new cases. The authors selected the most infected cities in India, and their findings suggested that there would be more cases in the coming months [15].

As of July 2020, there were 15,947,292 laboratory-confirmed cases and 642,814 deaths worldwide. India has reported 1,338,928 confirmed cases and 31,412 deaths [16]. Ribeiro (2020) discussed many aspects of COVID-19, and an ARIMA time series forecasting model was used to predict the cases over the following 50 days. The findings predicted an upward trend, enabling the government to take the necessary protective measures [17]. The ARIMA (1,2,0) model predicted COVID-19 deaths over the two months after the study, showing that 75,000 people might have been infected by the middle of September 2020 [18].

ARIMA and double exponential smoothing were used to predict future cases in Algeria. Data from March to November 2020 were collected by the Algerian Ministry of Health. The ARIMA (0,1,1) was used to forecast future cases in a given period. The results generated by this model were accurate after validation of the model. The projected COVID-19 cases confirmed that the recovered cases and deaths followed an exact pattern during the three days examined [19].

Smoothed data and independent variables were also used to improve a model's accuracy. Future research is needed to improve the ARIMA model's accuracy in forecasting COVID-19 cases. The best model is that with the lowest values of the performance metrics [20]. Ceylan (2020) used data from February to August 2020 to predict the next 30 days of cases. The findings revealed that the results generated by the multi-layer perceptron (MLP) network and the Holt-Winters model were accurate. Approximately 2,500 cases and 100 deaths were predicted to occur on September 14, 2020. According to the findings, some models were unable to predict future cases [21]. There will be new cases, but deaths will remain consistent or decrease based on predicted outcomes. Furthermore, COVID-19 prevention measures will help public health bodies and the government develop the appropriate policies to control the impact of the pandemic in Iran. Proper monitoring and precautionary measures will play a significant role in controlling the pandemic. Many studies have proposed time series–based models to predict upcoming cases [22]. The Holt-Winters, Prophet, LSTM, and ARIMA models can accurately predict future cases. A study used the ARIMA time series forecasting model to forecast the future cumulative cases in Spain, Italy, and France using data from February to April 2020. They analyzed many ARIMA model parameter values for p, q, and d, then used ARIMA (0,2,1) to accurately predict future cases. These findings may assist governments and health sectors in drafting appropriate policies [23].

Six models were used to forecast future cases in Switzerland, Turkey, Belgium, Germany, the United Kingdom, Finland, France, and Denmark. The findings found that the long short-term memory (LSTM) model was more accurate than others in forecasting future cases. In the second study, the model predicted the subsequent 14 days of cases. According to these results, the cumulative future case growth rate in many countries was expected to decrease. COVID-19 analysis and prediction are challenging due to changes in time series data [24]. Three different techniques were used to analyze COVID-19 time series data, revealing that the results generated by the LSTM model were more accurate than those of ARIMA and the nonlinear autoregression neural network (NARNN). From February 15, 2020, to June 2020, the online database collected daily time series data on total confirmed cases from the five top countries. The ARIMA time series model was used to forecast the active cases for the next 77 days [25].

The model's accuracy was cross validated using evaluation metrics—mean absolute percentage error (MAPE), median absolute percentage error (MdAPE), mean squared error (MSE), and RMSE. The forecast graph showed a slight increase in future cases for Russia and Spain, whereas the United States, Brazil, and India presented an exponential trend. Their findings predicted that 14 lakh and 25 lakh people from India and Brazil, respectively, would be infected by the end of July 2021. In contrast, 4.3 million people were predicted to be infected in the United States. Because no effective cure currently exists, this forecast will help the government and the healthcare sector increase healthcare facilities to reduce future confirmed and recovered cases [26]. Many articles have proposed using mathematics and time series models to forecast the upcoming pandemic's impact. The autoregressive moving average (ARMA) was used to forecast future cases based on existing data from Saudi Arabia. The authors used all combinations of ARIMA parameters to determine the best model. The findings showed that the ARIMA model produces better results than others [27]. The findings further showed an exponential trend in the predicted cases. Therefore, strict preventive policies should be enforced to slow the spread of the virus in Saudi Arabia, or there will be approximately 7,000 cases per day. Time series–based models were proposed to predict future cumulative deaths and active cases in the 16 most infected countries. Alzahrani (2020) selected the ARIMA model parameters using the auto.arima function in R and found that the seasonal ARIMA (SARIMA) model predicted future cases more accurately than the ARIMA model [28].

3. Theory/calculations

3.1. Time series analysis

Time series data are comprised of data values recorded over time (e.g., daily, weekly, monthly, or yearly). Time series can be either stationary or non-stationary. Stationary time series are those that present no specific pattern. Stationarity plays a vital role in time series analysis. Therefore, differencing and logging techniques are used to make non-stationary time series into stationary time series. First- and second-order differencing are given in equations 1 and 2, respectively. Second-order differencing is always followed by first-order differencing.

$$S_t^{'} = S_t - S_{t-1} \tag{1}$$

$$S_t^{''} = S_t - 2S_{t-1} + S_{t-2} \tag{2}$$

where S' is first-order differencing, S'' is second-order differencing, S_{t-1} is the observation at the (t−1) timestamp, and S_{t-2} is the observation at the (t−2) timestamp.

Time series forecasting requires model training. Each model has parameters that sense trends and cycles (seasonality) in time series data. The main objective of time series analysis is to find a model that can appropriately describe the patterns and predict future outcomes. COVID-19 data is recorded over time; therefore, we can analyze this data using time series models. This study considers data from January 2020 to August 2021.

3.2. Autoregressive integrated moving average (ARIMA [p,d,q])

To use the ARIMA model to estimate future cases, we first ensured no trends or seasonality were present by checking for stationarity. The time series data must have no upward or downward trend or seasonality for it to be considered stationary with constant mean-variance.

***Trend*:** When the data include a significant increase or decrease.

***Seasonality*:** When a time series has a repeated pattern over a given period (e.g., year, month, or day)

Time series: $S_1 \cdots S_n$ when S_i——S_j points are placed in a fixed pattern.

Lag: The lag of a given time series can be defined as its ith lag. S_t is the observation of the time series, S_{t-i}.

$$\mathbf{L}(\mathbf{S}_t) = \mathbf{S}_{t-1} \tag{3}$$

3.2.1. Hyperparameters of the ARIMA model

Autoregressive model (p): The autoregressive (AR) model of time series can be represented by a linear function with some noise or error in its previous values. It is also known as a memory-based model.

$$s_t = c + \varnothing_1 s_{t-1} \ldots + \varnothing_p s_{t-p} + \varepsilon_t \tag{4}$$

where S_t = time series, c = intercept constant, φ_i = coefficient that measures the impact of the initial values on the value of S_t, and ε_i = univariate white noise.

The moving average model (q): The moving average (MA) model of time series can be represented by a linear function with some univariate white noise of the last q+1 random shock, which is generated by ε_i:

$$s_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} \ldots + \theta_q \varepsilon_{t-q} \tag{5}$$

Autoregressive moving average model (p, q): The ARMA model of time series can be represented by the summation of the AR and MA models:

$$s_t = c + \varnothing_1 s_{t-1} \ldots + \varnothing_p s_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} \ldots + \theta_q \varepsilon_{t-q} \tag{6}$$

Differencing (d): Differencing is a technique for reducing or removing trends and seasonality, making time series stationary. Differencing is performed according to requirements. If we subtract the current value, the previous value is called the first-order difference transform. For example, X_t time series can be created by first-order differencing.

$$Y_t = \mathbf{X}_t – \mathbf{X}_{t-1} \tag{7}$$

3.3. Analytical tools and model evaluation

3.3.1. Autocorrelation function (ACF) and partial autocorrelation function (PACF)

Autocorrelation is the correlation between the present observed value and the previously observed value. An autocorrelation plot with lags is called an autocorrelation plot. An ACF shows the linear relationship between the observations at time t and previous observations at time t−n. The ACF for a given time series X can be defined as:

$$\boldsymbol{ACF}(\boldsymbol{X_t},\ \boldsymbol{X_{t-n}}) = \frac{\boldsymbol{Covariance}(\boldsymbol{X_t},\ \boldsymbol{X_{t-n}}\ )}{\boldsymbol{Variance}(\boldsymbol{X_t})} \tag{8}$$

where n is the lag (or difference between X_t and $X_{t\text{-}n}$).

In the PACF plot between observed values X_t and X_{t-2}, n = 2 can be defined as:

$$\boldsymbol{PACF}(\boldsymbol{X_t}, \boldsymbol{X_{t-2}}) = \frac{\boldsymbol{Covariance}(\boldsymbol{X_t}, \boldsymbol{X_{t-2}}/\boldsymbol{X_{t-1}})}{\sqrt{\boldsymbol{variance}(\boldsymbol{X_t}/\boldsymbol{X_{t-1}})}\sqrt{\boldsymbol{variance}(\boldsymbol{X_{t-2}}/\boldsymbol{X_{t-1}})}} \tag{9}$$

3.3.2. Akaike information criterion and bayesian information criterion

The Akaike information criterion (AIC) and Bayesian information criterion (BIC) are information criteria to check the goodness of the proposed model. This information helps assess the model's parameters and how well the model performed. Both functions set the lower value to determine which model can achieve the highest likelihood value.

$$\boldsymbol{AIC} = -2\boldsymbol{logL}(\widehat{\boldsymbol{\theta}}\ ) + 2\boldsymbol{n} \tag{10}$$

$$\boldsymbol{BIC} = -2\boldsymbol{logL}(\widehat{\boldsymbol{\theta}}\ ) + \boldsymbol{nlogN} \tag{11}$$

where log L $(\widehat{\theta})$ is the likelihood function, n is the number of model parameters, and N is the number of observations.

3.3.3. Evaluation metrics

Evaluation metrics were used to assess the proposed model's accuracy:

$$\boldsymbol{MAE} = \frac{1}{\boldsymbol{p}}\sum_{t=1}^{p}\left|\boldsymbol{S_t} - \boldsymbol{S_t'}\right| \tag{12}$$

$$\boldsymbol{MSE}\ = \frac{1}{\boldsymbol{p}}\sum_{t=1}^{P}\left(\boldsymbol{S_t} - \boldsymbol{S_t'}\right)^2 \tag{13}$$

$$\boldsymbol{RMSE} = \sqrt{\frac{1}{\boldsymbol{p}}\sum_{t=1}^{p}\left(\boldsymbol{S_t} - \boldsymbol{S_t'}\right)^2} \tag{14}$$

$$\boldsymbol{MAPE} = \frac{1}{\boldsymbol{p}}\sum_{t=1}^{p}\left|\frac{\boldsymbol{S_t} - \boldsymbol{S_t'}}{\boldsymbol{S_t}}\right| \tag{15}$$

where $\mathbf{S_t}$ is the actual value and S_t' is the predicted value.

A logarithmic approach may be necessary to make the time series stationary after differencing. This approach takes the log value of each point, followed by differencing.

3.3.3.1. Augmented Dickey–Fuller test. The augmented Dickey–Fuller (ADF) test determines whether the time series is stationary. If the p-value is lower than 0.05, the null hypothesis (H0) is rejected and the given time series is stationary. If the p-value is 0.05 or greater, we fail to reject H0, and multi-order differencing and logarithmic scaling are required to make the time series stationary.

- Null hypothesis (H0): the time series is not stationary and has a unit root.
- Alternative hypothesis (H1): the time series is stationary and has no unit root values.

3.3.3.2. ARIMA (AR, MA, I). The appropriate AR (p), MA (q), and I (d) values were determined using the following iterative process:

Step 1. Stationarity testing

Test the stationarity of the time series before applying the ARIMA model.

Step 2. Differencing

If the given time series has upward or downward trends and seasonality, then perform first-order differencing and check that the time series is stationary. Perform differencing according to the requirements to make the mean and variance constant.

Step 3. Determine the best parameters

Identify the optimal parameters using auto.arima and select the models based on the information criteria (AIC, BIC).

Step 4. Choose AR/MA/I

Develop the selected model based on the ACF and PACF plots of the residuals.

Step 5. Create the model

Apply the proposed model to predict the future occurrence by giving the period parameter.

Step 6. Test the model

Validate the accuracy of the proposed model by comparing the forecast values with the actual values.

The implementation of the ARIMA model is available at Google Colab, here.

3.4. Facebook prophet time series model

Prophet is a powerful and fast open-source time series model developed by Facebook using the C++ programming language. It uses an additive regression model to fit nonlinear trends with seasonality and holiday effects. The Prophet model uses the Fourier order for yearly seasonality, but weekly, then dummy variables are used. Prophet requires a minimum of two columns (y and ds), where ds is the time stamp and y is the value. The model is represented by:

$$\boldsymbol{S_t} = \boldsymbol{L_t} + \boldsymbol{Y_t} + \boldsymbol{H_t} + \boldsymbol{\varepsilon_t} \tag{16}$$

where S_t is the time series, L_t is the logistic/linear growth curve for fitting nonlinear changes, Y_t is seasonality, H_t denotes holiday effects, and ε_t denotes errors due to unusual changes.

3.4.1. The trend model (L_t)

This study implemented two trend models (the saturated growth and piecewise linear models) that cover all time series–related applications. The saturated growth/nonlinear model uses a logistic growth model to determine the trend of the given time series, represented by:

$$\boldsymbol{L_t} = \frac{\boldsymbol{N}}{1 + \mathbf{exp}(-\boldsymbol{k}(\boldsymbol{t} - \boldsymbol{m}))} \tag{17}$$

where N is the carrying capacity, k is the growth rate, and m is the offset parameter.

If there are continuous changes in the capacity, then the above equation cannot be used. In this case, the equation must be modified to capture the continuous changes in capacity with constant to varying capacity over the period N_t.

If there are C changepoints over C_i time, i.e., 1, 2, 3, … i, then a rare adjustment vector can be defined as:

$$\boldsymbol{\delta} \in \boldsymbol{R^C} \tag{18}$$

where δ_i is the change in rate that occurs at time C_i.

The rate for any given period is then the base rate R plus the adjustment rate for that period:

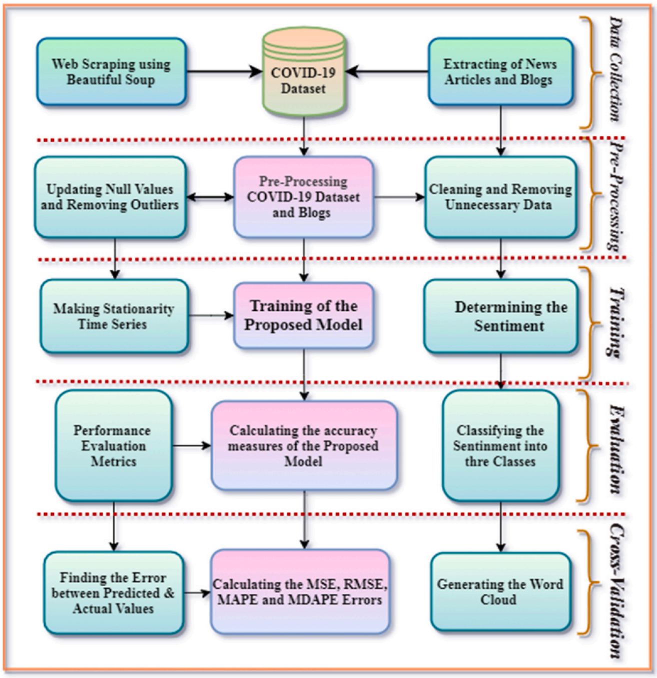
Web Scraping using Beautiful Soup
COVID-19 Dataset
Extracting of News Articles and Blogs
Data Collection
Updating Null Values and Removing Outliers
Pre-Processing COVID-19 Dataset and Blogs
Cleaning and Removing Unnecessary Data
Pre-Processing
Making Stationarity Time Series
Training of the Proposed Model
Determining the Sentiment
Training
Performance Evaluation Metrics
Calculating the accuracy measures of the Proposed Model
Classifying the Sentiment into thre Classes
Evaluation
Finding the Error between Predicted & Actual Values
Calculating the MSE, RMSE, MAPE and MDAPE Errors
Generating the Word Cloud
Cross-Validation

Fig. 3. The proposed methodology.

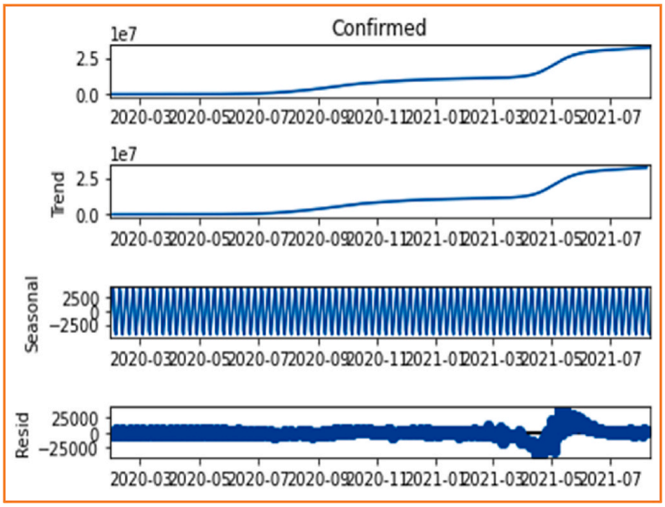

Fig. 4. Trend, seasonality, and residual of confirmed cases.

$$R + \sum_{i:t>C_i} \delta_i \tag{19}$$

$$a(t) \in \{0,1\}^C \tag{20}$$

$$a_i(t) = \left\{ \begin{matrix} 1, & if\ t > C_i, \\ 0, & otherwise \end{matrix} \right\} \tag{21}$$

The correct adjustment at any of the changepoints and the vector values defined above can be captured by:

$$\gamma_i = \left(C_i - m - \sum_{j<i} \gamma_j \right) \left(1 - \frac{k + \sum_{j<i} \delta_j}{k + \sum_{j\le i} \delta_j} \right) \tag{22}$$

The piecewise trend model (growth = logistic) is then given as:

$$L(t) = \frac{N(t)}{1 + \mathbf{exp}\left(-k + a(t)^T \delta \right) \left(t - \left(m + a(t)^T \gamma \right) \right)} \tag{23}$$

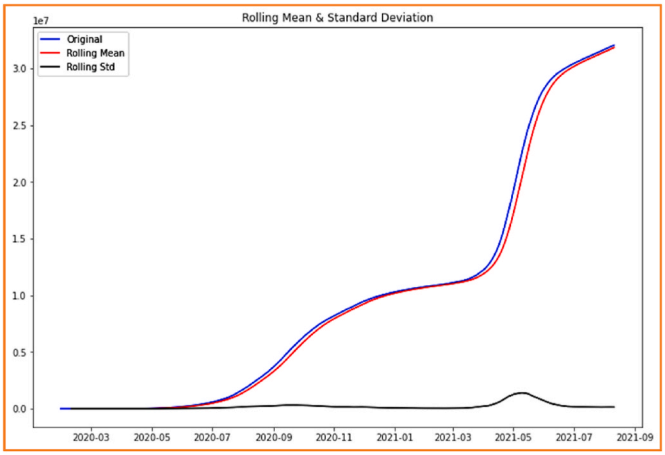

Fig. 5. Rolling mean and standard deviation of confirmed cases.

Table 3
ADF test and after log parameter values.

Parameters	Values	After Log	After 1st-Order Differencing
Test Statistic	0.898771	−3.151168	−2.592978
p-value	**0.993074**	**0.022991**	**0.094457**
# Lags Used	17.000000	16.000000	19.000000
No. of Observations	542.000000	542.000000	538.000000
Critical Value (1%)	−3.442473	−3.442473	−3.442563
Critical Value (5%)	−2.866887	−2.866887	−2.866927
Critical Value (10%)	−2.569618	−2.569618	−2.569639

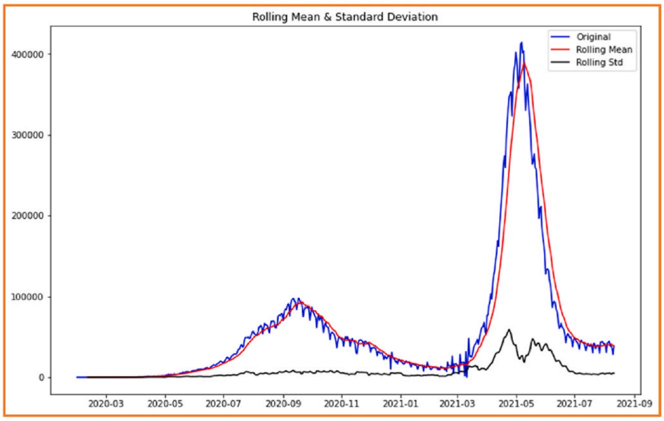

Fig. 6. Rolling mean and standard deviation after log.

3.4.2. *Seasonality hyperparameter*

Seasonality plays a considerable role in adopting periodic changes in time series data. Seasonality can be daily, weekly, or yearly. For example, a business's data for weekdays and weekends will differ if its clients buy more products during weekends than on weekdays. In the case of COVID-19 data, infection rates have been high from March to June every year. Prophet uses the Fourier order to model the seasonality. P is a periodic value: yearly = 365.25, weekly = 7, and daily = 1.

$$Y(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi nt}{P} \right) \right) + b_n \sin\left(\frac{2\pi nt}{P} \right) \tag{24}$$

The seasonality mode hyperparameter is the primary hyperparameter for the seasonality, indicating how much the seasonality component is integrated into the prediction. The default seasonality mode is additive, which is used for constant seasonality and trends;

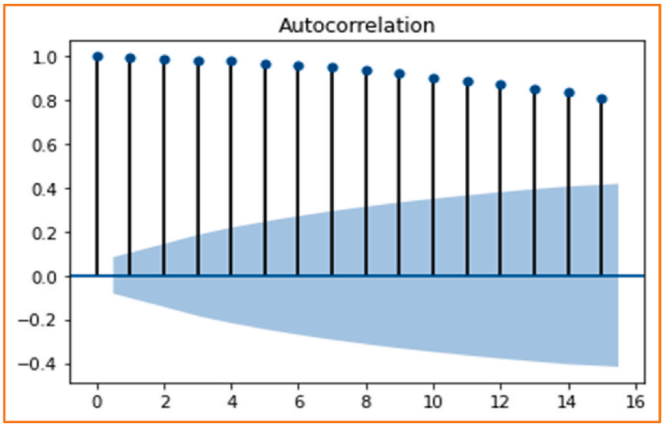

Fig. 7. Acf plot of confirmed cases.

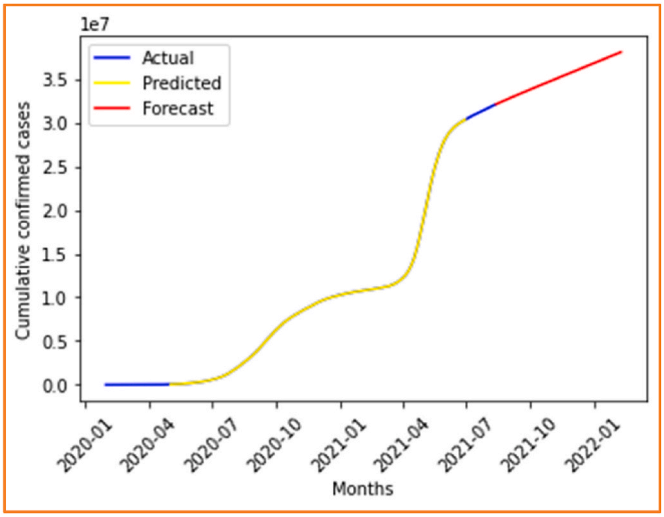

Fig. 8. Forecasting by the proposed ARIMA (cumulative confirmed cases).

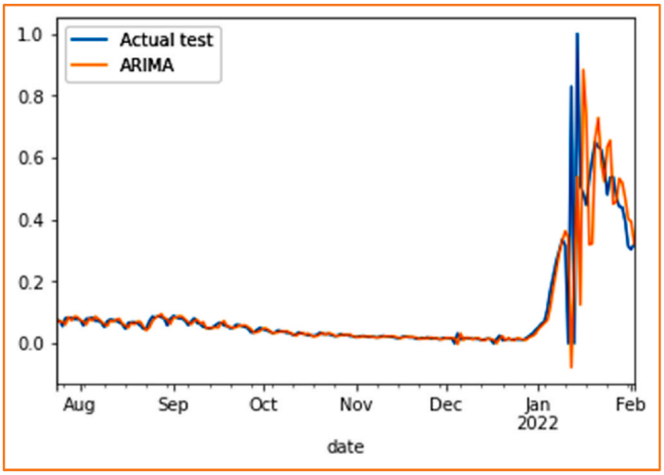

Fig. 9. Forecasting by the proposed ARIMA (daily confirmed cases).

otherwise, the mode should be multiplicative. The seasonality prior scale hyperparameter allows for flexible seasonality.

3.4.3. *Growth hyperparameter*

Growth is the most specific parameter to understand and implement when we know the data. The growth parameter value is set as either linear or logistic. When we plot the data and see rising trends with no saturation insight, we set the growth parameter to linear; otherwise, it is set to logistic, in which case we must provide the data's minimum and maximum reach to the prediction and actual data.

3.4.4. *Holiday hyperparameter*

Holidays result in days, weeks, or months affecting the time series analysis. In the case of forecasting future COVID-19 cases, there tend to be more cases on weekends because more people go outside during the weekend than on weekdays in India. These days must be considered in the model using the holiday parameter. The holiday parameter was necessary because daily COVID-19 data were used to forecast daily cases. Another holiday hyperparameter that deals with the effect of holidays on the prediction is the holidays prior scale.

3.4.5. *Changepoint hyperparameter*

Changepoints are another model hyperparameter and consider changes in the trend. For instance, in April 2021, India had the most COVID-19 cases of any country. There are four types of changepoint hyperparameters: changepoint prior scale, n changepoints, changepoint range, and changepoints. When changepoint dates are provided to the model, the model will not discover any more changepoints. Therefore, the model was allowed to discover the changepoints on its own and set the number of changepoints using the n changepoints hyperparameter for better results. The number of changepoints depends on each particular use case. For the COVID-19 use case, we set one changepoint every week.

The changepoint prior scale determines the flexibility of the particular changepoint that is allowed (how much fits the data). When this value is too high, overfitting occurs. The changepoint range does not affect the performance of the model as considerably as other hyperparameters. Thus, it was left at its default value for better results.

3.4.6. *Bias and variance in time series analysis*

Bias and variance are part of the model's reducible error and are essential parameters for building an accurate model. The reducible error requires the appropriate selection of the model so that its complexity and flexibility can be managed during the model's training.

Bias, also known as "error due to squared bias," is the difference between the predicted and targeted classes during the model's training. The resampling technique used the appropriate predicted value to achieve more accurate results (desired bias) to reduce the difference between the actual and predicted values. Therefore, in the time series analysis, bias can affect the overall prediction of the model.

Underfitting occurs when the error is high and overfitting occurs when the error is much lower than the predicted and targeted values during the model's training. Underfitting results in high accuracy during the training phase and low accuracy during the testing phase and is marked by high bias and variance, whereas overfitting is characterized by low bias and high variance. Bias is the error of the training phase, and variance is the error of the testing phase. Generally, bias and variance should both be low.

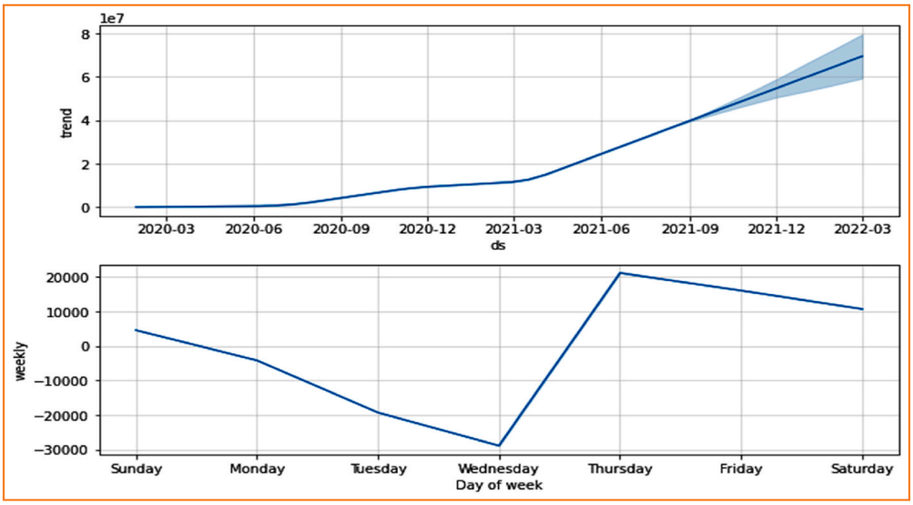

Fig. 10. Overall (monthly and weekly) trend.

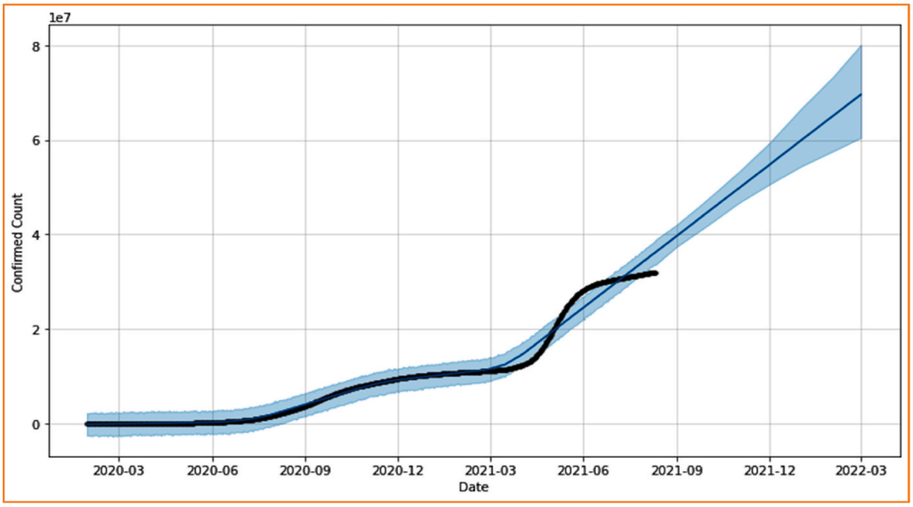

Fig. 11. Prediction by proposed prophet model.

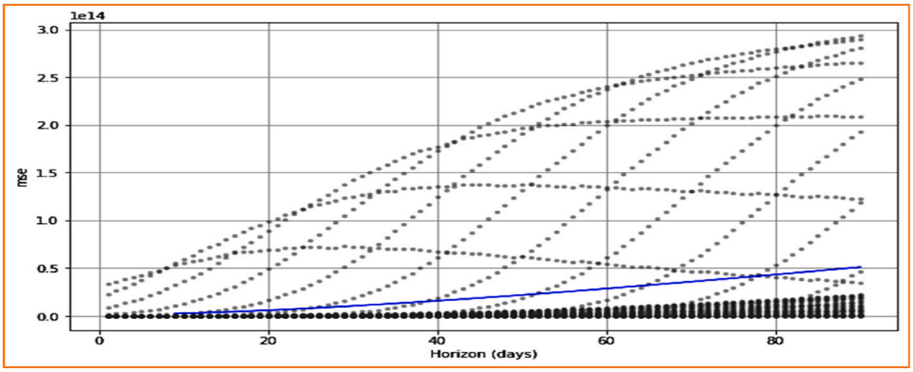

Fig. 12. Cross validation plot (MSE).

Table 4
SARIMAX results of the proposed ARIMA (1,2,2).

Dep. Variable:	Confirmed	No. Observations:	560
Model:	ARIMA (1,2,2)	Log-Likelihood	−5,779.855
Date:	Sat, September 11, 2021	AIC	11,567.710
Time:	18:17:47	BIC	11,585.008
Sample:	01-30-2020 to 08-11-2021	HQIC	11,574.465
Covariance Type:	opg		

3.5. Sentiment analysis

To learn the opinions of the health experts, scientists, and virologists regarding upcoming waves of the COVID-19 pandemic, we extracted related articles and news stories via web scraping with the Python package Beautiful Soup, which is used to parse HTML and XML web pages and extract keywords. Approximately 200 articles were extracted with Beautiful Soup as well as manually. Natural language processing (NLP) libraries were used to determine the sentiments from the extracted dataset.

Table 5
Coefficients of the proposed ARIMA (1,2,2).

	coef	std err	z	P > |z|	[0.025	0.975]
ar. L1	0.9615	0.011	88.563	0.000	0.940	0.983
ma. L1	−1.0398	0.019	−53.939	0.000	−1.078	−1.002
ma. L2	0.1576	0.021	7.355	0.000	0.116	0.200
sigma2	6.337e+07	6.81e-11	9.31e+17	0.000	6.34e+07	6.34e+07

Table 6
Summary of the proposed ARIMA (1,2,2).

Ljung Box (L1) (Q):	**0.55**	Jarque Bera (JB):	**1,177.83**
Prob (Q):	0.46	**Prob (JB):**	0.00
Heteroscedasticity (H):	483.22	**Skew:**	−0.85
Prob (H) (two-sided):	0.00	**Kurtosis:**	10.43

Table 7
Initial prediction by Prophet.

trend	yhat lower	yhat upper	trend lower	trend upper	yhat
6.499576e+07	5.619166e+07	7.290385e+07	5.639183e+07	7.257264e+07	6.497644e+07
6.962686e+07	5.832973e+07	8.029341e+07	5.925532e+07	7.896211e+07	6.960754e+07
7.475415e+07	6.141334e+07	8.675533e+07	6.202543e+07	8.690896e+07	7.477018e+07
7.971604e+07	6.430167e+07	9.485449e+07	6.436557e+07	9.495638e+07	7.972060e+07
8.484333e+07	6.621694e+07	1.022151e+08	6.646909e+07	1.025106e+08	8.481441e+07

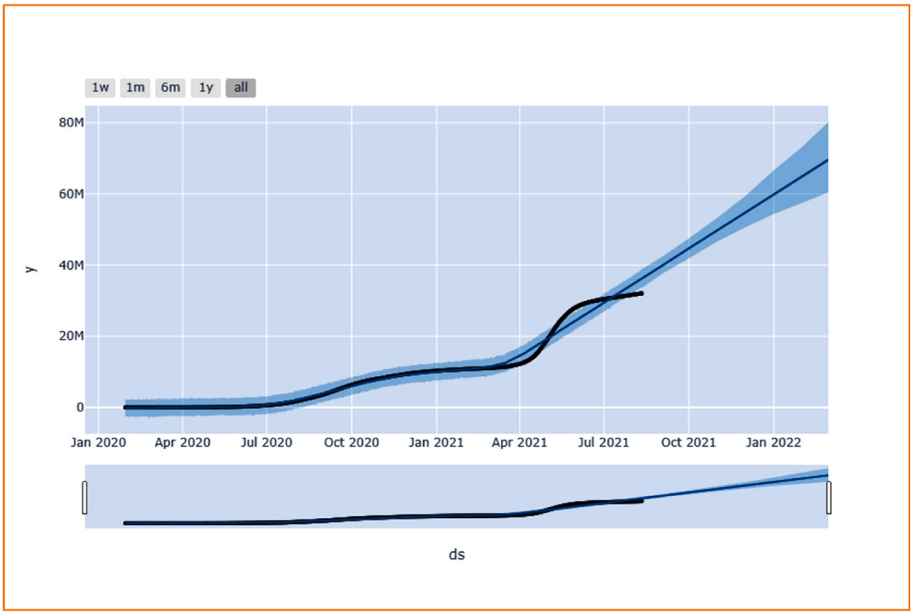

Fig. 13. Forecasting by the proposed prophet (cumulative confirmed cases).

Table 8
Cross-Validation of the Prophet model from March 10, 2020, to August 11, 2021.

ds	yhat	yhat lower	yhat upper	y	cutoff
2020–03–10	17.105278	0.716286	34.057535	58	2020-03-09
2020–03–11	21.504746	4.469184	38.730697	60	2020-03-09
2020–03–12	22.921459	6.439937	40.124224	74	2020-03-09
2020–03–13	23.088273	6.289238	41.124438	81	2020-03-09
2020–03–14	23.755008	6.365515	40.079387	84	2020-03-09
2021–08–07	2.577697e+07	2.245349e+07	2.889221e+07	31895385	2021-05-13
2021–08–08	2.587800e+07	2.271847e+07	2.925299e+07	31934455	2021-05-13
2021–08–09	2.597733e+07	2.256805e+07	2.945662e+07	31969954	2021-05-13
2021–08–10	2.607027e+07	2.258676e+07	2.949502e+07	31998158	2021-05-13
2021–08–11	2.616913e+07	2.261766e+07	2.970208e+07	32036511	2021-05-13

Table 9
Proposed prophet model performance metrics (diagnostics).

horizon	mse	rmse	mae	mape	mdape	coverage
9 days	2.491663e+12	1.578500e+06	622397.820887	0.146576	0.067136	0.040404
10 days	2.748039e+12	1.657721e+06	662582.775243	0.154969	0.073924	0.037879
11 days	3.017092e+12	1.736978e+06	703421.226818	0.163270	0.079725	0.037879
12 days	3.298150e+12	1.816081e+06	744934.722980	0.171243	0.087448	0.037879
13 days	3.593368e+12	1.895618e+06	787285.835527	0.179066	0.093846	0.037879

Table 10
Final prediction (prophet).

	ds	yhat	yhat lower	yhat upper
562	2021-11-01	4.977510e+07	4.667466e+07	5.331550e+07
563	2021-12-01	5.471227e+07	5.058744e+07	5.938837e+07
564	2022-01-01	5.987917e+07	5.438893e+07	6.664042e+07
565	2022-02-01	6.497644e+07	5.754582e+07	7.329056e+07
566	2022-03-01	6.960754e+07	6.045268e+07	8.018420e+07

3.5.1. *Removing unnecessary metadata from the dataset*

Before processing the scraped dataset for sentiment analysis, it was necessary to remove some extraneous data to achieve accurate results: e. g., special characters, URLs, # hashtags, and stop words. We used the Texthero Python library to clean the dataset.

3.5.2. *Determining sentiment using TextBlob*

After removing the unnecessary metadata from the dataset, the TextBlob Python package was used to determine the sentiments.

Table 11
Comparative studies between state of the art and proposed model (Prophet and ARIMA).

State of the art models				Proposed study country (India)				
Model	Country	Metrics	Values	Horizon	MSE	RMSE	MAPE	MDAPE
ML RF [38]	Worldwide	MAE	368.82	09 days	02.49	01.57	00.145	00.067
ML KNN [37]	India	MAE	649.74	10 days	02.74	01.65	00.154	00.073
ARIMA [41]	India	MAE	47.42	11 days	03.01	01.73	00.163	00.079
ML RF [34]	India	RMSE	717.73	12 days	03.29	01.81	00.171	00.087
DL LSTM [43]	USA	RMSE	324.61	13 days	03.59	01.89	00.179	00.093
DL LSTM [42]	Worldwide	RMSE	307.58	14 days	03.87	01.94	00.186	00.101
Holt Winter [37]	India	MAE	269.39	15 days	04.16	02.01	00.192	00.107
ARIMA [40]	Spain	RMSE	379.89	16 days	04.41	02.09	00.201	00.115
SARIMA [37]	India	RMSE	98.717	17 days	04.67	02.17	00.208	00.122
GBR [37]	India	RMSE	678.74	18 days	04.94	02.24	00.208	00.129

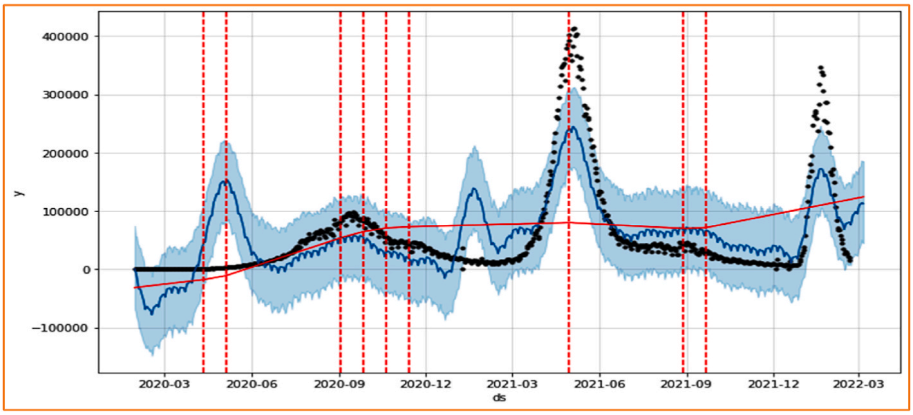

Fig. 14. Forecasting by the proposed prophet (daily confirmed cases).

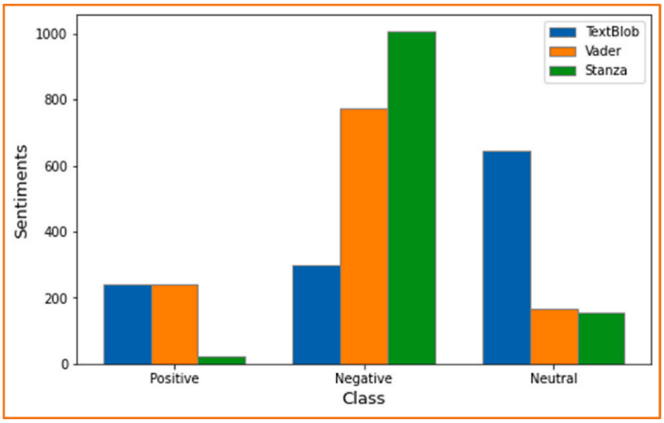

Fig. 15. Comparison of NLP libraries.

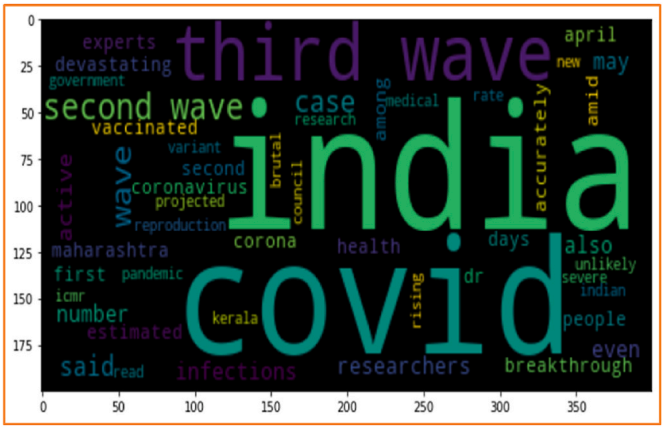

Fig. 16. Word cloud for negative sentiments.

TextBlob categorizes the sentiments by subjectivity and polarity score. Positive polarity indicates positive sentiment, negative polarity indicates negative sentiment, and a polarity of 0 indicates neutral sentiment.

3.5.3. *Determining sentiment using VADER*

NLTK (the Natural Language Toolkit) is an open-source Python library for NLP. It provides pre-trained models that are primarily used for processing textual data. VADER is a rule-/lexicon-based fully open-source library protected under the MIT license, mainly designed to analyze social media text using a bag of words approach. VADER returns a single unidimensional score with a range of −1 to +1. Positive values mean the sentiment of a given text is positive, and negative values mean sentiment is negative. Neutral sentiment values are between −0.05 and + 0.05.

3.5.4. *Determining sentiment using stanza*

Stanza is an NLP library created by the Stanford NLP Group to analyze text in more than 70 languages. It uses the CNN classifier model to classify the sentiment of a given dataset. It produces a score from 0 to 2, where 0 represents the negative class, 1 represents the neutral class, and 2 represents the positive class. We chose the mean value of each sentence because each paragraph may contain many sentences. The proposed methodology is given in Fig. 3.

4. Results and discussion

4.1. *Prediction by autoregressive integrated moving average (ARIMA)*

We checked the stationarity of the time series of confirmed cases before applying the ARIMA model by examining the p-value to reject H0. Fig. 4 revealed some properties of the confirmed cases to help determine stationarity. We further checked the stationarity with the ADF test. After applying the rolling statistics and the ADF test. Fig. 5 and Table 3 present the characteristics of the time series data and show that the p-value (0.99) is not lower than the threshold value of 0.05. We therefore cannot reject the null hypothesis H0, which means the times series has a unit root; hence, it is not stationary. We applied the log approach and conducted the ADF test again to ensure stationarity of the time series data. The "values after log" are shown in Table 3 and Fig. 6. The p-value (0.022) is lower than the threshold value after using the log. Thus, the given time series is stationary and the null hypothesis can be rejected. The ARIMA model was then applied to the time series to find the optimal parameters for accurate prediction results. After applying auto.arima and plotting the ACF and PACF (Fig. 7), the model with the optimal parameters was determined to be ARIMA (1,2,2).

Fig. 8 presents the actual dataset (blue), predicted dataset (yellow), and forecasted dataset (red). This study used COVID-19 data from January 2020 to August 2021 [44]. We trained the model (predicted values) on the data from May 2020 to May 2020. The data were forecast for the following 180 days.

Fig. 8 shows that ARIMA (1,2,2) can predict future cumulative confirmed cases. Prediction of daily confirmed cases is given Fig. 9. The results of the model's cross-validation are given in Tables 4–6. ARIMA (1,2,2) refers to AR (p) = 1, MR (q) = 2, and differencing (d) = 2. Log-Likelihood represents the maximum likelihood estimation. The AIC results from the model parameters and maximum likelihood values and helps to evaluate the strength of the model. Both the AIC and BIC values aid in feature selection and determining the model's reliability (see Fig. 11) (see Fig. 12) (see Fig. 10).

Table 5 shows the significance of each feature. The row ar. L1 represents autoregression with a lag of one, and ma. L1 and ma. L2 represent moving averages with a lag of one and two, respectively. The std err column shows the estimation of the error of the predicted value and the strength of the effect of the residual error on the estimated parameters. The standardized coefficient (z) values are coef and standard error. If these values exceed the threshold (0.05), the predicted values may be unreliable. The current model parameters were deemed acceptable because the p-value is less than 0.05. The last two columns in Table 5 show the confidence intervals with marginal error.

The Ljung Box (L1) (Q) shown in Table 6 tested for the absence of serial autocorrelation (white noise) at a lag of 1. Heteroscedasticity (H) tested for error residuals with the same variance or different variance. The summary of the model shows a heteroscedasticity (H) of 483.22 and a probability (H) of 0.00, which is lower than the threshold value. Hence, we can reject the null H0 hypothesis. The residuals show some variance. Jarque Bera (JB) tests the normality of error and null distributions against the alternative of another distribution. The JB is 1,177.83, and Prob (JB) is 0.0, which means we can reject the null hypothesis as the data is not normally distributed.

4.2. *Prediction by facebook prophet*

After setting the future date as March 1, 2022, we applied the Prophet model to predict the yhat values (future cases) based on past data. Table 7 shows the predicted (yhat) and validated (yhat) values based on the lower and upper bounds of yhat. The trend values were validated by trend lower and trend upper; the yhat values must be between yhat lower and yhat upper. Prediction of the future cumulative confirmed cases is given in Fig. 13.

Table 8 shows how the time series data were fitted to Prophet with cutoff dates between March 9, 2020, and May 13, 2021, to cross-validate the generated values with an initial of 30 days and a period of 10 horizons, a maximum of 90 days. The results (yhat) generated by the model are between the range of yhat lower and yhat upper. We then fit the future created dates onto the current model to predict the cumulative future cases. Prediction of the future daily confirmed cases is given in Fig. 14 (see Table 10) (see Table 11) (see Table 9).

5. Conclusion

This study's main aim was to predict the future daily confirmed and cumulative confirmed cases of the third wave of COVID-19 in India. The ARIMA and Prophet time series forecasting models were used to predict the future daily confirmed and cumulative confirmed cases. NLP libraries (TextBlob, VADER, and Stanza) were used for sentiment analysis. The results show that both models can predict future cases based on past cases. However, the Prophet model is better than the ARIMA model for long-term forecasting. There will likely be more cases in the third wave because the proposed model shows an exponential curve. However, deaths and recovered cases might be affected by factors like new variants, herd immunity, vaccinations, and resource availability. In the second wave in India, the Delta-1 variant was more infectious and deadly than the other COVID-19 variants. Over 55% of India's eligible adult population is now fully vaccinated against COVID-19, and 172 crore vaccine doses have been administered. This vaccination rate will play a significant role in the third wave. India could achieve herd immunity through vaccination and indirectly from the second wave because every third person was infected with the virus. However, a new COVID-19 variant might be challenging for public health authorities and governments. A comparison between different sentiment libraries is given in Fig. 15, and a word cloud for negative sentiment is given in Fig. 16. This study does not consider transmissibility and other factors while making the predictions. All materials and the implemented model's python code are available at the GitHub Repository here.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

No funding to declare; Cambridge (https://proofreading.org/) for English language editing and proofreading.

References

[1] W.H. Organization, WHO Announces COVID-19 Outbreak a Pandemic, 2020, https://doi.org/10.4324/9781003095590-6. Available from: https://www.euro.who.int/en/health-topics/healthemergencies/coronavirus-COVID-19/news/news/2020/3/who-announcesCOVID-19-outbreak-a-pandemic.

[2] H.D. Meares, M.P. Jones, When a system breaks: queueing theory model of intensive care bed needs during the COVID-19 pandemic, Med. J. Aust. 212 (10) (2020 Jun) 470–471, https://doi.org/10.5694/mja2.50605. Epub 2020 May 7. PMID: 32379951; PMCID: PMC7267520.

[3] D.S. Gaur, Global forecasting OF COVID-19 using arima based prophet, Intern. J. Eng. Appl. Sci. Technol. 5 (2) (2020) 463–467, https://doi.org/10.33564/ijeast.2020.v05i02.077.

[4] T.T. Tran, L.T. Pham, Q.X. Ngo, Forecasting epidemic spread of SARS-CoV-2 using ARIMA model (Case study: Iran), Global J. Soil Sci. Environ. Manag. 6 (2020) 1–10, https://doi.org/10.22034/GJESM.2019.06.SI.01. Special Issue (COVID-19).

[5] H. Abbasimehr, R. Paki, Prediction of COVID-19 confirmed cases combining deep learning methods and Bayesian optimization, Chaos, Solit. Fractals 142 (2021) 110511, https://doi.org/10.1016/j.chaos.2020.110511.

[6] S.P. Marbaniang, Forecasting the Prevalence of COVID-19 in Maharashtra, Delhi, Kerala, and India using an ARIMA model. https://doi.org/10.21203/rs.3.rs-34555/v1, 2020.

[7] G. Perone, An ARIMA Model to Forecast the Spread and the Final Size of COVID-2019 Epidemic in Italy, 2020, https://doi.org/10.1101/2020.04.27.20081539 medRxiv.

[8] S. Ghosal, S. Sengupta, M. Majumder, B. Sinha, Linear Regression Analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases-March 14th 2020), Diabetes, Metab. Syndrome: Clin. Res. Rev. 14 (4) (2020) 311–315, https://doi.org/10.1016/j.dsx.2020.03.017.

[9] P. Furtado, Epidemiology SIR with regression, arima, and Prophet in forecasting COVID-19, Eng. Proceed. 5 (No. 1) (2021) 52, https://doi.org/10.3390/engproc2021005052. Multidisciplinary Digital Publishing Institute.

[10] B. Fanoodi, B. Malmir, F.F. Jahantigh, Reducing demand uncertainty in the platelet supply chain through artificial neural networks and ARIMA models, Comput. Biol. Med. 113 (2019) 103415, https://doi.org/10.1016/j.compbiomed.2019.103415.

[11] Hedi, M.V.J. Merawati BR, Modeling and forecasting of COVID-19 confirmed cases in Indonesia using ARIMA and exponential smoothing, in: Proceedings of the International Seminar of Science and Applied Technology, 2020, https://doi.org/10.2991/aer.k.201221.043. ISSAT 2020).

[12] L.R. Kundu, M.Z. Ferdous, U.S. Islam, M. Sultana, Forecasting the spread of COVID-19 pandemic in Bangladesh using ARIMA model, Asian J. Med. Biol. Res. 7 (1) (2021) 21–32, https://doi.org/10.1101/2020.10.22.20217414.

[13] D. Parbat, M. Chakraborty, A python based support vector regression model for prediction of COVID19 cases in India, Chaos, Solit. Fractals 138 (2020) 109942, https://doi.org/10.1016/j.chaos.2020.109942.

[14] L. Bayyurt, B. Bayyurt, Forecasting of COVID-19 Cases and Deaths Using ARIMA Models, 2020, https://doi.org/10.1101/2020.04.17.20069237 medrxiv.

[15] M. Panda, Application of ARIMA and Holt-Winters Forecasting Model to Predict the Spreading of COVID-19 for India and its States, 2020, https://doi.org/10.1101/2020.07.14.20153908.

[16] Jbrahim Sabry, Forecasting COVID-19 cases in Egypt using ARIMA-based time series analysis, Eurasian J. Med. Oncol. (2021), https://doi.org/10.14744/ejmo.2021.64251.

[17] M. Maleki, M.R. Mahmoudi, D. Wraith, K.H. Pho, Time series modelling to forecast the confirmed and recovered cases of COVID-19, Trav. Med. Infect. Dis. 37 (2020) 101742, https://doi.org/10.1016/j.tmaid.2020.101742.

[18] M.H.D.M. Ribeiro, R.G. da Silva, V.C. Mariani, L. dos Santos Coelho, Short-term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil, Chaos, Solit. Fractals 135 (2020) 109853, https://doi.org/10.1016/j.chaos.2020.109853.

[19] M. Abdelaziz, A. Ahmed, A. Riad, G. Abderrezak, A.-A. Djida, Forecasting daily confirmed COVID-19 cases in Algeria using ARIMA models. https://doi.org/10.1101/2020.12.18.20248340, 2020.

[20] N. Talkhi, N. Akhavan Fatemi, Z. Ataei, M. Jabbari Nooghabi, Modeling and forecasting number of confirmed and death caused COVID-19 in Iran: a comparison of time series forecasting methods, Biomed. Signal Process Control 66 (2021) 102494, https://doi.org/10.1016/j.bspc.2021.102494.

[21] Z. Ceylan, Estimation of COVID-19 prevalence in Italy, Spain, and France, Sci. Total Environ. 729 (2020) 138817, https://doi.org/10.1016/j.scitotenv.2020.138817.

[22] R. Salgotra, M. Gandomi, A.H. Gandomi, Time series analysis and forecast of the COVID-19 pandemic in India using genetic programming, Chaos, Solit. Fractals (2020) 109945, https://doi.org/10.1016/j.chaos.2020.109945.

[23] İ. Kırbaş, A. Sözen, A.D. Tuncer, F.Ş. Kazancıoğlu, Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches, Chaos, Solit. Fractals 138 (2020) 110015, https://doi.org/10.1016/j.chaos.2020.110015.

[24] A.K. Sahai, N. Rath, V. Sood, M.P. Singh, ARIMA modelling & forecasting of COVID-19 in top five affected countries, Diabetes, Metab. Syndrome: Clin. Res. Rev. 14 (5) (2020) 1419–1427, https://doi.org/10.1016/j.dsx.2020.07.042.

[25] V.K.R. Chimmula, L. Zhang, Time series forecasting of COVID-19 transmission in Canada using LSTM networks, Chaos, Solit. Fractals (2020) 109864, https://doi.org/10.1016/j.chaos.2020.109864.

[26] CDC, COVIDView Weekly Summary, 2020. Available from: https://www.cdc.gov/coronavirus/2019-ncov/COVID-data/COVIDview/index.html.

[27] S.F. Ardabili, A. Mosavi, P. Ghamisi, F. Ferdinand, A.R. Varkonyi-Koczy, U. Reuter, P.M. Atkinson, COVID-19 outbreak prediction with machine learning, Algorithms 13 (10) (2020) 249, https://doi.org/10.3390/a13100249.

[28] S.I. Alzahrani, I.A. Aljamaan, E.A. Al-Fakih, Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions, J. Infect. Publ. Health 13 (7) (2020) 914–919, https://doi.org/10.1016/j.jiph.2020.06.001.

[29] T. Ye, X. Yang, Analysis and prediction of confirmed COVID-19 cases in China with uncertain time series, Fuzzy Optim. Decis. Making 20 (2) (2021) 209–228, https://doi.org/10.1007/s10700-020-09339-4.

[30] K.E. ArunKumar, D.V. Kalaga, C.M. Sai Kumar, G. Chilkoor, M. Kawaji, T. M. Brenza, Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA), Appl. Soft Comput. 103 (2021) 107161, https://doi.org/10.1016/j.asoc.2021.107161.

[31] P.G. Jamdade, S.G. Jamdade, Modeling and prediction of COVID-19 spread in the Philippines by October 13, 2020, by using the VARMAX time series method with preventive measures, Results Phys. 20 (2021) 103694, https://doi.org/10.1016/j.rinp.2020.103694.

[32] E. Gecili, A. Ziady, R.D. Szczesniak, Forecasting COVID-19 confirmed cases, deaths and recoveries: revisiting established time series modeling through novel applications for the USA and Italy, PLoS One 16 (1) (2021), e0244173, https://doi.org/10.1371/journal.pone.0244173.

[33] S. Roy, G.S. Bhunia, P.K. Shit, Spatial prediction of COVID-19 epidemic using ARIMA techniques in India, Model. earth sys. environ. 7 (2) (2021) 1385–1391, https://doi.org/10.1007/s40808-020-00890-y.

[34] M. Zivkovic, N. Bacanin, K. Venkatachalam, A. Nayyar, A. Djordjevic, I. Strumberger, F. Al-Turjman, COVID-19 cases prediction by using hybrid machine learning and beetle antennae search approach, Sustain. Cities Soc. 66 (2021) 102669, https://doi.org/10.1016/j.scs.2020.102669.

[35] R. Salgotra, A.H. Gandomi, Time series analysis of the COVID-19 pandemic in Australia using genetic programming, Data Sci. COVID-19 (2021) 399–411, https://doi.org/10.1016/B978-0-12-824536-1.00036-8. Academic Press.

[36] J. Devaraj, R.M. Elavarasan, R. Pugazhendhi, G.M. Shafiullah, S. Ganesan, A. K. Jeysree, E. Hossain, Forecasting of COVID-19 cases using deep learning models: is it reliable and practically significant? Results Phys. 21 (2021) 103817, https://doi.org/10.1016/j.rinp.2021.103817.

[37] T. Saba, I. Abunadi, M.N. Shahzad, A.R. Khan, Machine Learning Techniques to Detect and Forecast the Daily Total COVID-19 Infected and Deaths Cases under Different Lockdown Types. Microscopy Research and Technique, 2021, https://doi.org/10.1002/jemt.23702.

[38] Z. Malki, E.S. Atlam, A. Ewis, G. Dagnew, A.R. Alzighaibi, G. ELmarhomy, I. Gad, ARIMA models for predicting the end of COVID-19 pandemic and the risk of second rebound, Neural Comput. Appl. 33 (7) (2021) 2929–2948, https://doi.org/10.1007/s00521-020-05434-0.

[39] R. Patil, U. Patel, T. Sarkar, COVID-19 cases prediction using regression and novel SSM model for non-converged countries, J. Appl. Sci. Eng. Technol. Edu. 3 (1) (2021) 74–81, https://doi.org/10.35877/454RI.asci137.

[40] L. Bayyurt, B. Bayyurt, Forecasting of COVID-19 cases and deaths using ARIMA models, medRxiv (2020), https://doi.org/10.1101/2020.04.17.20069237.

[41] R. Anne, ARIMA Modelling of Predicting COVID-19 Infections, 2020, https://doi.org/10.1101/2020.04.18.20070631 medRxiv.

[42] C. Direkoglu, M. Sah, Worldwide and regional forecasting of coronavirus (COVID-19) spread using a deep learning model. https://doi.org/10.1101/2020.05.23.20111039, 2020.

[43] Y. Tian, I. Luthra, X. Zhang, Forecasting COVID-19 Cases Using Machine Learning Models, MedRxiv, 2020, https://doi.org/10.1101/2020.07.02.20145474.

[44] Covid19, Covid19 India, 2021. https://www.covid19india.org. (Accessed 11 August 2021).