SN Computer Science (2021) 2:416
https://doi.org/10.1007/s42979-021-00808-0

ORIGINAL RESEARCH

Multiplicative Holts Winter Model for Trend Analysis and Forecasting of COVID-19 Spread in India

H. Swapnarekha[1] · Himansu Sekhar Behera[1] · Janmenjoy Nayak[2] · Bighnaraj Naik[3] · P. Suresh Kumar[4]

Received: 4 June 2020 / Accepted: 3 August 2021 / Published online: 16 August 2021

Abstract

The surge of the novel COVID-19 caused a tremendous effect on the health and life of the people resulting in more than 4.4 million confirmed cases in 213 countries of the world as of May 14, 2020. In India, the number of cases is constantly increasing since the first case reported on January 30, 2020, resulting in a total of 81,997 cases including 2649 deaths as of May 14, 2020. To assist the government and healthcare sector in preventing the transmission of disease, it is necessary to predict the future confirmed cases. To predict the dynamics of COVID-19 cases, in this paper, we project the forecast of COVID-19 for five most affected states of India such as Maharashtra, Tamil Nadu, Delhi, Gujarat, and Andhra Pradesh using the real-time data. Using Holt–Winters method, a forecast of the number of confirmed cases in these states has been generated. Further, the performance of the method has been determined using RMSE, MSE, MAPE, MAE and compared with other standard algorithms. The analysis shows that the proposed Holt–Winters model generates RMSE value of 76.0, 338.4, 141.5, 425.9, 1991.5 for Andhra Pradesh, Maharashtra, Gujarat, Delhi and Tamil Nadu, which results in more accurate predictions over Holt's Linear, Auto-regression (AR), Moving Average (MA) and Autoregressive Integrated Moving Average (ARIMA) model. These estimations may further assist the government in employing strong policies and strategies for enhancing healthcare support all over India.

Keywords COVID-19 · Holt–Winters · Holt's linear · Auto regression · Moving average · ARIMA

Introduction

Throughout history, it is evident that different contagious diseases have claimed the lives of many people and caused difficult conditions that take a long period to conquer the situation. In the past, the surge of smallpox has killed roughly 500 million people all over the world [1]. In 1918, an approximate of 17–100 million individuals has been killed due to the epidemic of Spanish influenza [2]. Several pandemics have been emerging from the last 20 years like severe acute respiratory syndrome coronavirus (SARS-CoV) in the year 2002–2003, H1N1 influenza in the year 2009, and the Middle East respiratory syndrome coronavirus (MERS-CoV) in the year 2015. The

This article is part of the topical collection "Computer Aided Methods to Combat COVID-19 Pandemic" guest edited by David Clifton, Matthew Brown, Yuan-Ting Zhang and Tapabrata Chakraborty.

✉ H. Swapnarekha
swapnarekha23@gmail.com

Himansu Sekhar Behera
hsbehera_india@yahoo.com

Janmenjoy Nayak
mailforjnayak@gmail.com

Bighnaraj Naik
mailtobnaik@gmail.com

P. Suresh Kumar
reshu.suri@gmail.com

[1] Department of Information Technology, Veer Surendra Sai University of Technology (VSSUT), Burla, Sambalpur 768018, Odisha, India

[2] Department of Computer Science and Engineering, Aditya Institute of Technology and Management (AITAM), Tekkali, Andhra Pradesh 532201, India

[3] Department of Computer Application, Veer Surendra Sai University of Technology, Burla, Sambalpur 768018, Odisha, India

[4] Department of Computer Science and Engineering, Dr. Lankapalli Bullayya College of Engineering (W), Visakhapatnam 530013, India

outbreak of novel coronavirus since December 2019 in the city of Wuhan in South China has killed above hundreds and infected more than thousands of individuals within the first few days of the pandemic. The human coronaviruses that have originated from the animal reservoirs in the twenty-first century lead to a global epidemic with frightening morbidity and mortality. These viruses are named corona due to the appearance of a spike-like morphology on the external area under the electronic microscope. It is composed of single-stranded RNA belonging to the Coronavirinae subfamily, which belongs to Coronaviridae family. α, β, γ and δ are the four genera of these viruses. Mammals are usually infected by α- and β-CoV, while the birds are infected by γ- and δ-CoV. Less pathogenicity and mild respiratory syndrome as the common cold are caused by the HCoV-229E and HCoV-NL63 of alpha coronavirus and HCoV-HKU1 and HCoV-OC43 of beta-coronavirus. While, severe and malignant breathing infections are exhibited by the SARS-CoV and MERS-CoV of β-CoVs [3].

In December 2019, local hospitals in City of Wuhan in South China were reported with people diagnosed with unidentified pneumonia [4]. All the people diagnosed with unidentified pneumonia were connected to the Huanan Seafood Market where varieties of live species are available. The symptoms of these cases are similar to the clinical characteristics of pneumonia caused by virus. On 7 January 2020, the Centers for disease control (CDC) experts after analyzing samples gathered from the throat swabs, declared the disease as novel coronavirus pneumonia (NCP) [5]. Later, the ICTV (International Committee on Taxonomy of Viruses) named the novel virus as SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) [6, 7]. On 11 February 2020, the World Health Organization (WHO) declared the disease as novel COVID-19 [8]. The COVID-19 induced by SARS-CoV-2 associates to β-CoV. The genome structure of SARS-CoV-2 exhibits 79.5% similarity to SARS-CoV as it sustains eight residues of the SARS-CoV-binding residues [9]. As the SARS-CoV-2 genome sequencing exhibits 96.2% similarity to Bat Coronavirus RaTG13, both the bat coronavirus and human SARS-CoV-2 use the similar ancestor [10]. On 30 January 2020, the WHO announced the surge as a Public Health Emergency of International Concern (PHEIC) after the dissemination of COVID-19 to 18 countries as a result of person-to-person contact. In the United States, the major crisis was established when they identified the first case that was not carried from China on 26 February 2020. When the number of COVID-19 infections has raised 13 times in different regions of the world other than China and when the number of countries affected by COVID-19 has tripled, then the WHO has announced COVID-19 as pandemic on 11 March 2020 as it causes serious threat to the public health all over the world. The number of COVID-19 cases registered in different countries of the world has crossed all the previous records of other pandemics over time. It is considered the most dangerous disease till date due to its rapid transmission [11].

The first COVID-19 case in India was registered on January 30, 2020. In the month of March, the total of COVID-19 infections started escalating. Most of these cases are connected to the people having travel history to other countries that are affected by COVID-19 [12]. The Indian government has implemented strict actions by suspending all visas to India with effect from 13 March 2020. As of May 14, 2020, the cumulative number of registered cases in India is 81,997. The number of daily registered COVID-19 cases in India up to May 14, 2020 over 7 days MA is shown in the following Fig. 1.

From the past decades, the progress in sensor technology, biological understanding, and mathematical techniques is contributing to the growing significance of modeling in the field of health and bioinformatics. A mathematical model can be described as a depiction of a system utilizing mathematical notions and language to facilitate appropriate interpretation of a system or to analyze the influence of different elements and to generate predictions on patterns of behavior [13]. As mathematical modeling activity insists transparency and certainty regarding inferences, it enables us to evaluate our understandings of the epidemiology of infection by correlating model results with the recognized patterns. In the field of medicine, mathematical models are suitable for performing research on epidemiology, planning, and assessment of precautionary and control programs, clinical investigations, health and cost–benefit analysis, investigation of patients and in maximizing the efficacy of operations directed in attaining stated goals with existing resources [14]. A statistical model incorporates a set of statistical assumptions to approximate reality and to make predictions from these approximations. The advantage of statistical models is that it summarizes the results of a test and presents them in such a way so that one can more easily see and understand any patterns within the data. The usage of a statistical model allows clinical analysts to obtain moderate and accurate assumptions from gathered information and to make reliable decisions in the existence of ambiguity.

A mathematical procedure known as decomposition method has been suggested by Adomian [15] to provide solutions to the problems of neuroscience, such as the conduction of nerve impulses, analyzing the behavior of the immune system or observation of medication effects, and so on. Further, the results demonstrate the accuracy and efficacy of the proposed method. A mathematical model to

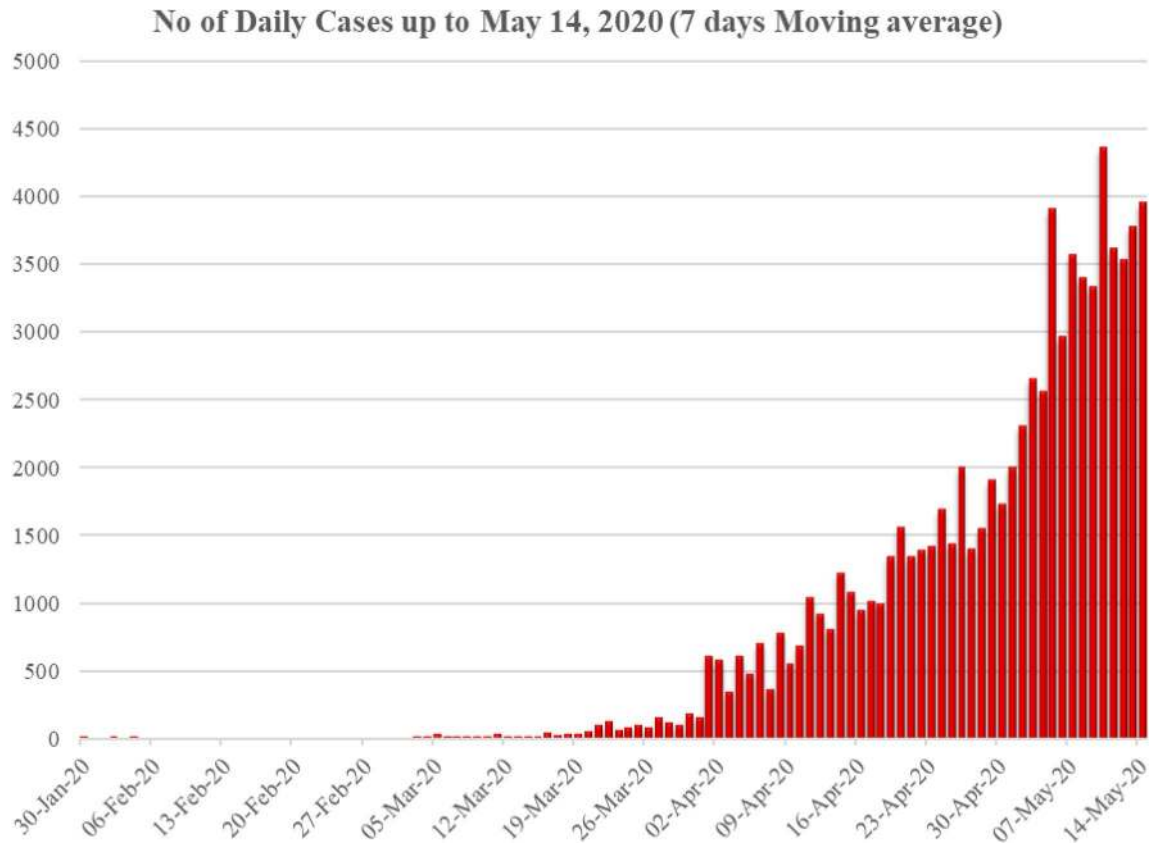

Fig. 1 No of daily cases registered in India up to May 14, 2020

predict whether isolation and quarantine can stop the spread of SARS has been developed by Castillo-Chavez et al. [16]. The amount of data required to predict SARS has been reduced due to the simplicity of questions and assumptions in the proposed model. Further, results indicate that the recommended model can reduce the size of the SARS outbreak by a factor of 1000. To determine the risk of non-immune persons obtaining dengue when traveling, a mathematical model has been represented by Massad et al. [17]. Further, the model is tested using Singapore data and the results depict the robustness of the proposed mathematical model in predicting the risk of getting dengue when traveling to countries having dengue-endemic. To forecast the spread of infectious diseases like dengue, two statistical models, namely ARIMA model and the Knorr–Held two-component (K–H) model, have been suggested by Earnest et al. [18]. The proposed models have been validated on Singapore dengue fever data. Further, the performance of the models has been distinguished with the Mean Absolute Percentage Error (MAPE). The results show that the K–H model results in a lesser MAPE value of 17.21 and takes a longer time to execute when compared to the ARIMA model. To analyze clinical data and more complicated data, the concept of linear and logistic regressions along with a modern statistical model known as Bayesian networks has been described by Yoo et al. [19]. Using the modern statistical model, the interactions among clinical, genomic, and environmental data have been represented. Further, it is also concluded that the modern statistical model outperforms in analyzing both clinical and complicated data. To analyze tuberculosis epidemiology, a statistical model named a Bayesian model has been proposed by Getoor et al. [20]. Statistical relation models which are constructed using a data-driven method are used to model distributions over relational domains. The model has been applied to the San Francisco tuberculosis patient data. Further, results indicate the potentiality of the proposed model over other conventional statistical approaches.

From the past few pandemics, the assessment of human loss and the prediction of mortality rate until certain period or closure of the pandemic has been performed successfully using the statistical models. In the present pandemic, researchers and technocrats have been using the same statistical procedures in the assessment of spread rate and mortality rate as these models show better performance in the prediction of earlier epidemics. The statistical model based on multivariate analysis has been proposed by Xu et al. [21] to determine the false-negative results as well as window period for testing positive. This model is used to determine the clinical symptoms that are important for detecting the false-negative results of SARS-CoV-2. Moreover, a prediction model based on the clinical characteristics has been proposed to identify the right time for testing. Further, the findings show that the proposed model provides better accuracy in the clinical diagnosis of the COVID-19 pandemic. To estimate the dynamics of disease transmission over time, a statistical model combined with data of COVID-19 cases in Wuhan has been proposed by Kucharski et al. [22]. The proposed model has been evaluated on publicly available datasets on cases in Wuhan as well as on the International cases exported from Wuhan. Based on the findings, the authors concluded that there will be a decline in the transmission of COVID-19 in Wuhan during late January 2020. An analysis based on Boltzmann's function to predict the number of deaths in China has been proposed by Gao et al. [23]. From the findings, it can be concluded that the assessment of the severity of the situation can be better predicted using the proposed method. To calculate the real number of contaminated people and to assume the infection fatality ratio (IFR), a novel mechanistic statistical model combined with the SIR (Susceptible, Infected and Recovered) has been proposed by Roques et al. [24]. The findings show that the IFR is compatible with the earlier findings in China (0.66%) and lesser than the earlier computed value on the Diamond Princess Cruise ship data (1.3%). A statistical model based on Holt's second-order exponential smoothing method and ARIMA model has been proposed by Poonia and Azad [25] to forecast COVID-19 infected patients in 28 states and 5 union territories of India. From the results, it can be observed that the cumulative number of cases in India will increase to 36,335.63 and simultaneously the mortality rate may increase to 1099.38 by 1 May 2020. The other analysis done on the applicability of mathematical and statistical models has been depicted in the following Table 1.

Besides the successful implementation of statistical models in the prognosis and forecasting of the COVID-19 pandemic, yet certain limitations exist. The Moving-Average model performs well with stationary data. This model does not consider the trend or seasonality of time series data. In the Auto-regressive model, the assumption of uncorrelated error is easily violated as the independent variables are time-lagged values for the dependent variable. With the ARIMA model, the long-term forecasting generates poor prediction results. Although ARIMA model is the mostly used model for forecasting the time series, there are certain limitations of the model. The limitations of the ARIMA model are: (i) it does not have automatic updating feature as in smoothing models. Due to this reason, the entire modeling process has to be repeated from the beginning whenever new data are available, (ii) the likeness of ARIMA model to solve complex real-world problem is not always adequate as ARIMA models cannot handle the non-linear patterns [26], (iii) it does not provide support for changes in the middle of the prediction phases [27]. Therefore, in this paper, we propose Holt's–Winter model for forecasting the time series data with seasonal and trend patterns. Holt–Winters method is a time-series forecasting method that is used to extract and interpret data and statistics and portray results to more precisely forecast the future trend based on past data.

Proposed Methodology

In the time series analysis, error trend seasonality forecast (ETS), ARIMA and Holt–Winters are the main classical models that have been widely used as predictors. Holt's–Winter is a statistical model also called as triple exponential smoothing model used for short-term forecasting with seasonal and trend patterns. In Holt–Winters model, components, such as level, trend, and the season, are necessary for forecasting. The value of these components ranges between 0 and 1. Based on the pattern of the season, Holt–Winters model is classified as an additive model and multiplicative model. The additive method is considered when the variations in the season are constant throughout the series, while the multiplicative method is considered when the variations in the season change relative to the level of series. If the seasonal effect is independent of the prevailing mean level of the time series, then Holt–Winters additive model is used. If the seasonal effect is dependent on the mean level of the time series, i.e., the seasonal variations rise with the rise in mean level of time series, then Holt–Winter multiplicative model is used [28]. In COVID-19 time series data, trends can be observed due to the repetition of certain patterns on regular intervals of time because of external factors like lockdown of country, mandatory social distancing,

Table 1 Applicability of mathematical and statistical models in the prognosis and forecasting of COVID-19 disease

Author and year	Dataset	Method	Comparison of performance metrics	Parameters used	Ref
Chen and Yu (Mar, 2020)	Daily official reports of the National Health Commission of the People's Republic of China. (http://www.nhc.gov.cn/xcs/yqfkdt/gzbd_index.shtml)	Second derivative model	$R^2 = 0.9778$	α = number of expected cases at the baseline, β = growth rate per day	[34]
Gupta et al. (Apr, 2020)	Time series data provided by John Hopkins University	SEIR (susceptible, exposed, infectious, recovered) and regression model	RMSLE for SEIR = 1.52, Regression = 1.75	β = infectious rate, σ = incubation rate, γ = recovery rate, ξ = rate at which recovered people become susceptible due to low immunity	[35]
Al-qaness et al. (Mar, 2020)	COVID-19 dataset from WHO website (https: //www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/)	FPASSA-ANFIS (Flower pollination algorithm using the Salp swarm algorithm—adaptive neuro-fuzzy inference system)	RMSE = 5799, MAE = 4271, MAPE = 4.79, RMSRE = 0.07, $R^2 = 0.9645$	Standard gamma = 1.5, Switch probability = 0.8, C2 ∈ [0, 1], C3 ∈ [0, 1]	[36]
Tandon et al. (Apr. 2020)	Official website of Johns Hopkins University (https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html)	ARIMA model	MAPE = 4.1	X_t = predicted number of confirmed COVID-19 cases, Z_t = residual term α1, α2, β1 and β2 are parameters	[37]
Chintalapudi et al. (Apr. 2020)	Official website of the Italian Health Ministry (http://www.salute.gov.it/nuovocoronavirus)	ARIMA model	Registered cases: RMSE = 514.74. MAE = 324.16, MAPE = 6.35 Recovered cases: RMSE = 186.65. MAE = 112.27 MAPE = 15.60	p = an Autoregressive referred to use of ancient values in model, 'd' = the difference degree of integrated I(d) component, 'q' = model error	[38]
Kumar et al. (Jan. 2020)	Data from Johns Hopkins Corona Virus Resource Center (https://coronavirus.jhu.edu/)	ARIMA and Richerd's Model	ARIMA model RMSE for incidence = 12.10, Mortality = 0.833, Recovered = 3.34	p = order of the autoregressive part, d = degree of first differencing involved, q = order of the MA part	[39]
Anee and Jeeva (Jan. 2020)	Data from Johns Hopkins university (https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html)	ARIMA model	RMSE = 121.439, MAE = 47.42, MAPE = 7.76	p = AR part of the model, d = integrated part of the mode, q = the MA parameter	[40]
Poonia and Azad (Apr. 2020)	India cumulative number of infected cases and deaths Worldometer website (online available: https://www.worldometers.info/coronavirus/country/india/) state-wise cumulative number of infected cases and deaths from (https://www.mygov.in/corona-data/covid19-statewise-status)	SutteARIMA model	MAPE = 0.036	p = AR part of the model, d = integrated part of the mode, q = the MA parameter, α and β are the smoothing parameters used	[25]

R^2 coefficient of determination; *RMSLE* root mean squared log error; *MAPE* mean absolute percentage error; *RMSRE* root mean squared relative error; *RMSE* root mean square error; *MAE* mean absolute error

quarantines, etc. Therefore, in this research, multiplicative method has been considered as the variations in COVID data are quite frequent. In this method, the seasonal components are communicated in relative terms, such as percentages, and the series are seasonally balanced by isolating through the seasonal component. Algorithm 1 represents the procedure of Holt–Winters model for COVID forecasting. The algorithm of Holt–Winters multiplicative model makes use of state space model to provide exponential smoothing that is similar to the statistical foundations used in the regression and Box/Jenkins methodology [28]. In Holt–Winters multiplicative model, the relationship to Holt–Winters multiplicative smoothing equations is revealed by providing equivalent exponential smoothing equations for the transition equations of level and trend. The observation equation represented as "y_t" is used to disclose the relationship between time series and state variables. The parameter "p_t" represents the level for time series, "q_t" represents the growth per period and "r_t" represents the seasonal factor. The error term is represented by "ε_t" which are independent of the past value of time series and state variables. The parameters "$\hat{p}_t$" "$\hat{q}_t$and$\hat{r}_t$" represent the Holt–Winters multiplicative smoothing equations. The parameter "m" is used to represent the

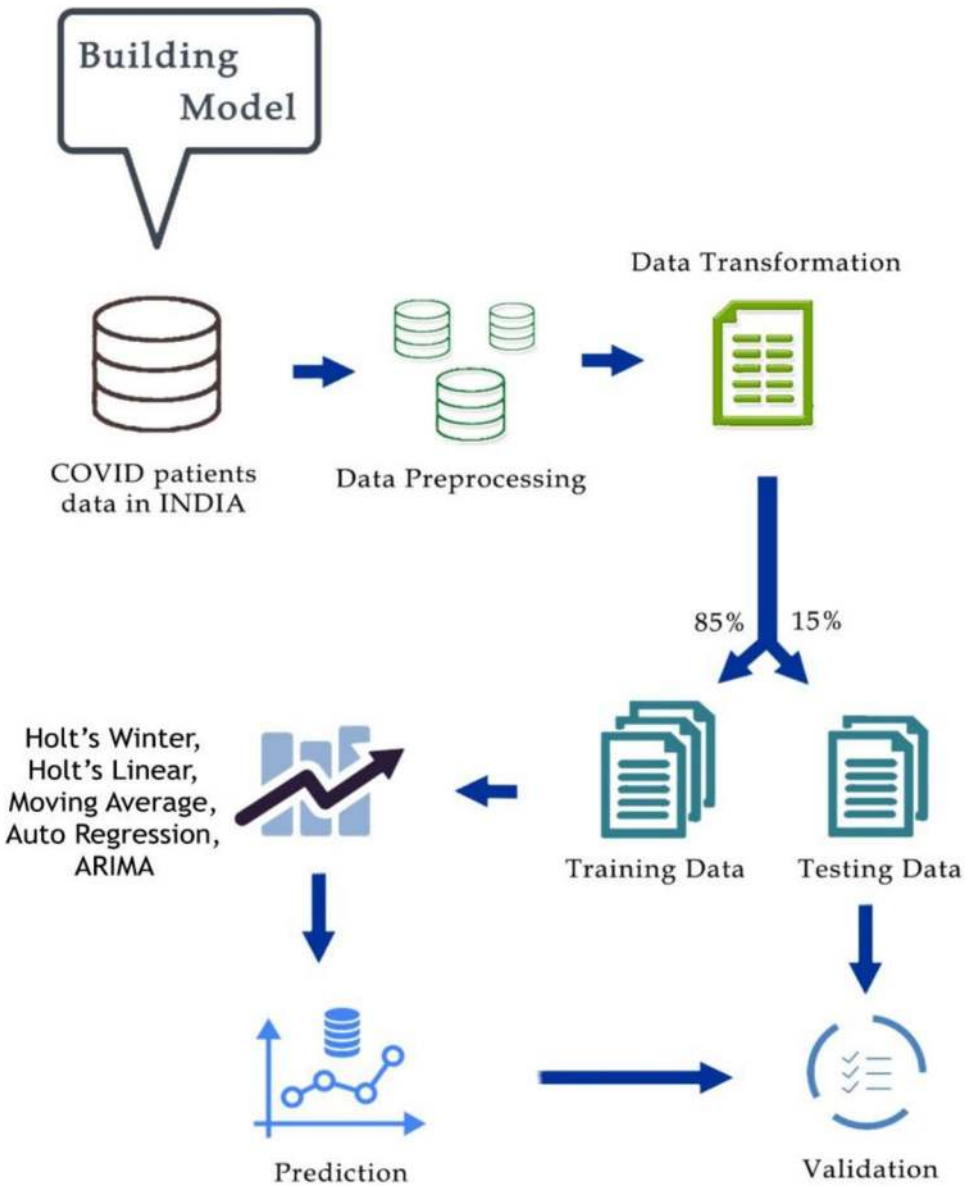

Fig. 2 Framework of the proposed method

frequency of seasonality that is the number of seasons in that particular year. The framework of the proposed work is represented in Fig. 2.

Algorithm 1: **Multiplicative Holt–Winters Model**

Step.1: Establish the relation between the time series y_t (Eq.1) and state variables (p_t , q_t , and r_t) at time t as presented in Eq.2, Eq.3, and Eq.4 respectively.

$$y_t = \left(p_{t-1} + q_{t-1}\right) \times r_{t-m} \times \left(1 + \varepsilon_t\right) = \left(p_{t-1} + q_{t-1}\right) \times r_{t-m} + \left(p_{t-1} + q_{t-1}\right) \times r_{t-m} \times \varepsilon_t \quad (1)$$

$$p_t = p_{t-1} + q_{t-1} + \beta_1 \times \left(p_{t-1} + q_{t-1}\right) \times \varepsilon_t \quad (2)$$

$$q_t = q_{t-1} + \beta_2 \times \left(p_{t-1} + q_{t-1}\right) \times \varepsilon_t \quad (3)$$

$$r_t = r_{t-m} + \beta_3 \times r_{t-m} \times \varepsilon \quad (4)$$

In Eq. 2 to Eq.4, β_1 , β_2 ,and β_3 are the smoothing constants in Holt–Winter's method. The $\beta_1, \beta_2, \beta_3 > 0$, $2 \times \beta_1 + \beta_2 < 4$, and $\beta_3 < 1$ are the necessary conditions for model's stability.

Step.2: Calculate the forecast error e_t (Eq.5) with perfect information at time $t-1$.

$$e_t = y_t - \left(p_{t-1} + q_{t-1}\right) \times r_{t-m} \quad (5)$$

Step.3: Compute the error term ε_t (Eq.6) from the given m number of seasons.

$$\varepsilon_t = \frac{y_t - \left(p_{t-1} + q_{t-1}\right) \times r_{t-m}}{\left(p_{t-1} + q_{t-1}\right) \times r_{t-m}} = \frac{e_t}{\left(p_{t-1} + q_{t-1}\right) \times r_{t-m}} \quad (6)$$

Step.4: Use $\{y_1, y_2 ... y_n\}$ to find estimates $\hat{p}_0$, $\hat{q}_0$, and $\hat{r}_{j-m}$ for p_0 , q_0 , and r_{j-m} .

Step.5: Find out the forecasts of h periods ahead at time n by using forecast function as in Eq.7, where $j = \left(h/m\right) + 1$.

$$\hat{y}(h) = \left(\hat{p}_n + h\hat{q}_n\right) \times \hat{r}_{n+h-jm} \quad (7)$$

Step.6: Update the estimates $\hat{p}_t$, $\hat{q}_t$, and $\hat{r}_t$ equations for $t = 1, 2, ..., n$ as in Eq.8, Eq.9 and Eq.10 respectively, where $\hat{e}_t = y_t - \hat{y}_t(1)$ and $\hat{y}_t(1) = \left(\hat{p}_{t-1} + \hat{q}_{t-1}\right) \times \hat{r}_{t-m}$.

$$\hat{p}_t = \hat{p}_{t-1} + \hat{q}_{t-1} + \hat{\beta}_1 \hat{e}_t / \hat{r}_{t-m} \quad (8)$$

$$\hat{q}_t = \hat{q}_{t-1} + \hat{\beta}_2 \hat{e}_t / \hat{r}_{t-m} \quad (9)$$

$$\hat{r}_t = \hat{r}_{t-m} + \hat{\beta}_3 \hat{e}_t / \left(\hat{p}_{t-1} + \hat{q}_{t-1}\right) \quad (10)$$

Experimental Setup

This research has been experimented on system setup with Lenovo T520 with Windows 10 Operating System and Intel Core i5 processor. The system is having 6 GB RAM. For experimentation, we have used python with packages, such as NumPy, pandas framework. Seaborn, Matplotlib modules are used for data visualization. For statistical analysis of the data in the models, statsmodel package has been used. In the study, proposed Holt–Winters model is compared with various statistical models, such as Holt's Linear, MA, AR and

Table 2 Parameter setting of various models with different states

State	Technique	Parameter setting
Andhra Pradesh	Holt–Winters Model	Seasonal_periods: 13 Trend: 'multiplicative' Seasonal: 'add'
	Holt's Linear Model	Smoothing_level: 1 Smoothing_slope: 0.4
	AR Model	p: 2 d: 2 q: 0
	MA Model	p: 3 d: 2 q: 0
	ARIMA	p: 1 d: 2 q: 0
Maharashtra	Holt–Winters Model	Seasonal_periods: 6 Trend: 'multiplicative' Seasonal: 'add'
	Holt's Linear Model	Smoothing_level: 1 Smoothing_slope: 0.4
	AR Model	p: 3 d: 1 q: 1
	MA Model	p: 1 d: 1 q: 3
	ARIMA	p: 1 d: 1 q: 1
Gujarat	Holt–Winters Model	Seasonal_periods: 23 Trend: 'add' Seasonal: 'add'
	Holt's Linear Model	Smoothing_level: 0.8 Smoothing_slope: 0.4
	AR Model	p: 3 d: 1 q: 3
	MA Model	p: 3 d: 1 q: 1
	ARIMA	p: 1 d: 1 q: 2
Delhi	Holt–Winters Model	Seasonal_periods: 46 Trend: 'mul' Seasonal: 'add'
	Holt's Linear Model	Smoothing_level: 0.3 Smoothing_slope: 0.4
	AR Model	p: 3 d: 1 q: 3
	MA Model	p: 2 d: 1 q: 2
	ARIMA	p: 1 d: 2 q: 0
Tamil Nadu	Holt–Winters Model	Seasonal_periods: 46 Trend: 'multiplicative' Seasonal: 'add'
	Holt's Linear Model	Smoothing_level: 1 Smoothing_slope: 2
	AR Model	p: 1 d: 0 q: 1
	MA Model	p: 4 d: 2 q: 2
	ARIMA	p: 3 d: 2 q: 0

ARIMA model. The data set is further divided into 85:15 ratio where 85 percent of data, i.e., the time period from 30 Jan 2020 to 28 Apr 2020 is used for training the model and 15 percent of data, i.e., the time period from 29 April 2020 to 14 May 2020 has been used for testing the model. The parameter setting of these models is displayed in Table 2.

Data Preprocessing

Feeding data and preparation of valid data are the primary steps in building a model. In this study, we considered the data of patients of different states in India from Covid19india.org [29] from January 30, 2020 to May 13, 2020. Data contain features, such as Date, Age Bracket, Gender, Patient_Status, City, District, State, State code, Notes, Nationality, Source_1, Source_2, Source_3. The data are imported as data frame using web scraping. Segregate the data frame with respective to date by considering the state and status, such as confirmed and recovered, divided the count into confirmed, deceased, and death. Finally, data are ready with columns Date, Name of State/UT, Latitude, Longitude, Total Confirmed cases, Death, Cured/Discharged/Migrated, New cases, New deaths, New recovered. Data are considered up to 14 May 2020. We considered the data with Name of State / UT as Maharashtra, Tamil Nadu, Delhi, Gujarat, and Andhra Pradesh and applied various models.

Performance Measures

In this section, we discussed the evaluation measure Root Mean Square Error (RMSE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE) which are used as a primary measure to evaluate the performance of the models. RMSE, MSE, MAPE, MAE are the standard metrics in regression and

Table 3 RMSE, MSE, MAE, and MAPE Scores of various models w.r.t. (a) Andhra Pradesh, (b) Maharashtra, (c) Gujarat, (d) Delhi, and (e) Tamil Nadu

(a) Andhra Pradesh				
Model name	RMSE	MSE	MAE	MAPE
Holt–Winters Model	76.0	5779.1	58.2	0.03
Holt's linear	162.2	26,293.8	123.9	0.06
AR Model	164.4	27,013.4	120.5	0.06
MA Model	172.3	29,704.2	127.5	0.06
ARIMA	147.4	21,720.2	106.3	0.05
(b) Maharashtra				
Model name	RMSE	MSE	MAE	MAPE
Holt–Winters Model	338.4	114,541.7	311.3	0.02
Holt's linear	2919.1	8,521,201.1	2444.3	0.11
AR Model	22,419.6	502,636,637.5	17,335.9	0.79
MA Model	12,997.4	168,932,968.1	9027.8	0.39
ARIMA	11,184.6	125,094,396.2	8343.5	0.37
(c) Gujarat				
Model name	RMSE	MSE	MAE	MAPE
Holt–Winters Model	141.5	20,023.9	126.6	0.02
Holt's linear	225.2	50,721.3	214.3	0.03
AR Model	11,349.8	128,817,078.2	8228.4	0.96
MA Model	8611.1	74,151,858.0	6139.5	0.71
ARIMA	6822.7	46,549,023.1	4834.8	0.56
(d) Delhi				
Model name	RMSE	MSE	MAE	MAPE
Holt–Winters Model	425.9	181,418.4	407.5	0.07
Holt's linear	1264.3	1,598,571.4	1132.1	0.17
AR Model	4931.2	24,316,280.7	3666.0	0.52
MA Model	5291.0	27,994,982.5	4003.7	0.58
ARIMA	486.7	236,835.8	462.6	0.08
(e) Tamil Nadu				
Model Name	RMSE	MSE	MAE	MAPE
Holt–Winters Model	1991.5	3,966,049.0	1696.3	0.25
Holt's linear	2554.1	6,523,389.1	2142.2	0.31
AR Model	4027.6	16,221,676.0	3437.0	0.50
MA Model	1491.9	2,225,829.6	1295.7	0.44
ARIMA	1378.6	1,900,604.2	1218.6	0.18

are helpful to know the efficiency of the model in terms of error rate. RMSE is calculated as the square root of the mean value of the squared difference between predictions and actual outcomes as shown in Eq. (11).

$$\text{RMSE} = \sqrt{\sum \frac{(\text{pred}_i - \text{actual}_i)^2}{\text{Total Predictions}}} \tag{11}$$

MSE is used to determine the average squared difference between the estimated and actual outcomes as shown in Eq. (12). The total number of predictions is indicated by 'n' in Eq. (12).

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\text{pred}_i - \text{actual}_i)^2 \tag{12}$$

Table 4 Forecast of total number of COVID-19 cases in Andhra Pradesh, Maharashtra, Gujarat, Delhi, and Tamil Nadu

State	Date	Total number of cases				
		Holt–Winters	Holt's linear	AR	MA	ARIMA
Andhra Pradesh	20/5/2020	2804.9	3177.6	3334.7	3367.4	3264.4
Maharashtra		39,075.8	25,440.5	188,944.8	156,536.3	124,213.9
Gujarat		11,610.3	11,098.6	179,751.7	137,001.3	110,623.3
Delhi		10,936.7	7061.7	41,735.7	43,173.7	11,720.8
Tamil Nadu		9203.2	6021.2	2283.6	11,359.7	13,651.0
Andhra Pradesh	26/5/2020	3457.8	4126.4	4640.3	4706.0	4499.1
Maharashtra		57,242.5	30,079.7	493,754.2	455,467.1	328,995.9
Gujarat		14,055.1	13,220.6	1,141,501.3	785,748.4	639,093.0
Delhi		15,300.4	8169.9	94,706.5	97,582.9	18,332.8
Tamil Nadu		13,644.1	7173.2	2126.6	17,807.0	26,622.9
Andhra Pradesh	1/6/2020	4191.9	5358.6	6640.2	6765.9	6370.7
Maharashtra		83,304.9	34,718.9	1,289,125.0	1,334,468.0	909,850.2
Gujarat		16,180.0	15,342.7	7,934,435.4	4,835,192.2	4,209,019.6
Delhi		21,118.9	9278.1	214,773.0	220,560.3	29,636.9
Tamil Nadu		19,753.2	8325.3	1983.1	26,925.8	54,752.2
Andhra Pradesh	7/6/2020	5103.6	6958.8	9771.7	10,007.6	9268.2
Maharashtra		120,694.5	39,358.1	3,362,884.1	3,917,301.7	2,591,854.5
Gujarat		18,584.0	17,464.7	57,794,525.8	30,684,471.7	29,849,608.8
Delhi		27,437.0	10,386.3	486,338.0	498,518.5	49,518.5
Tamil Nadu		27,073.6	9477.4	1851.7	39,240.2	118,741.2
Andhra Pradesh	14/6/2020	6356.2	9439.0	15,888.5	16,377.6	14,851.5
Maharashtra		185,313.8	44,770.5	10,282,469.5	13,768,481.6	9,014,888.4
Gujarat		21,350.5	19,940.4	603,534,665.3	267,992,376.3	308,901,807.5
Delhi		39,468.0	11,679.2	1,261,016.5	1,290,805.3	93,966.6
Tamil Nadu		40,421.1	10,821.4	1712.3	58,096.9	313,309.3
Andhra Pradesh	19/6/2020	7437.6	11,735.3	23,014.6	23,840.7	21,272.9
Maharashtra		251,287.2	48,636.5	22,831,415.2	33,795,801.3	22,211,256.6
Gujarat		23,374.6	21,708.8	3,256,255,430.4	1,276,999,961.7	1,671,168,247.1
Delhi		51,019.7	12,602.7	2,489,450.5	2,546,765.4	152,628.4
Tamil Nadu		54,370.6	11,781.5	1620.9	74,541.2	654,885.3

MAE is used to determine the errors among paired observations signifying the similar circumstance. In Eq. (13), 'n' indicates the total number of predictions

Table 5 Actual registered cases in Andhra Pradesh, Maharashtra, Gujarat, Delhi and Tamil Nadu

Date	Total number of cases				
	Andhra Pradesh	Maharash-tra	Gujarat	Delhi	Tamil Nadu
20/5/2020	2560	39,297	12,539	11,088	13,191
26/5/2020	2983	54,758	14,829	14,465	17,728
01/6/2020	3676	70,013	17,217	20,834	23,498
07/6/2020	4659	85,975	20,097	28,936	31,667
14/6/2020	6152	1,07,958	23,590	41,182	44,661
19/6/2020	7961	1,24,331	26,198	53,116	54,449

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}(\mathrm{pred}_i - \mathrm{actual}_i)}{n} \tag{13}$$

MAPE is standard loss function used to denote the prediction accuracy of forecasting as displayed in Eq. (14). The total number of predictions is represented using 'n' in the Eq. (14). The absolute value in this calculation is summed for every forecasted point in time and divided by the number of fitted points n.

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{(\mathrm{actual}_i - \mathrm{pred}_i)}{\mathrm{actual}_i}\right| \tag{14}$$

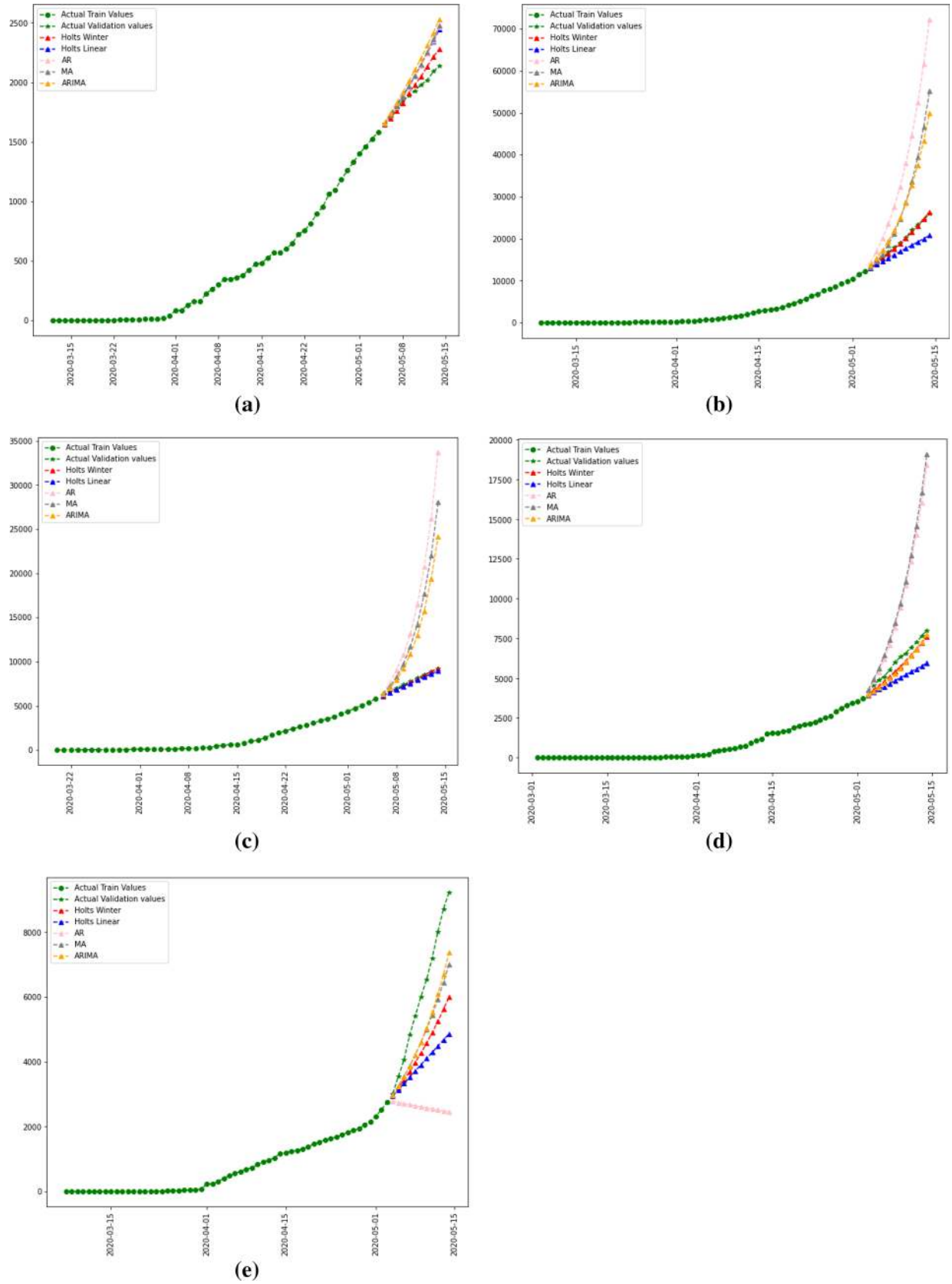

Fig. 3 Performance of various models **a** Andhra Pradesh, **b** Maharashtra, **c** Gujarat, **d** Delhi, and **e** Tamil Nadu

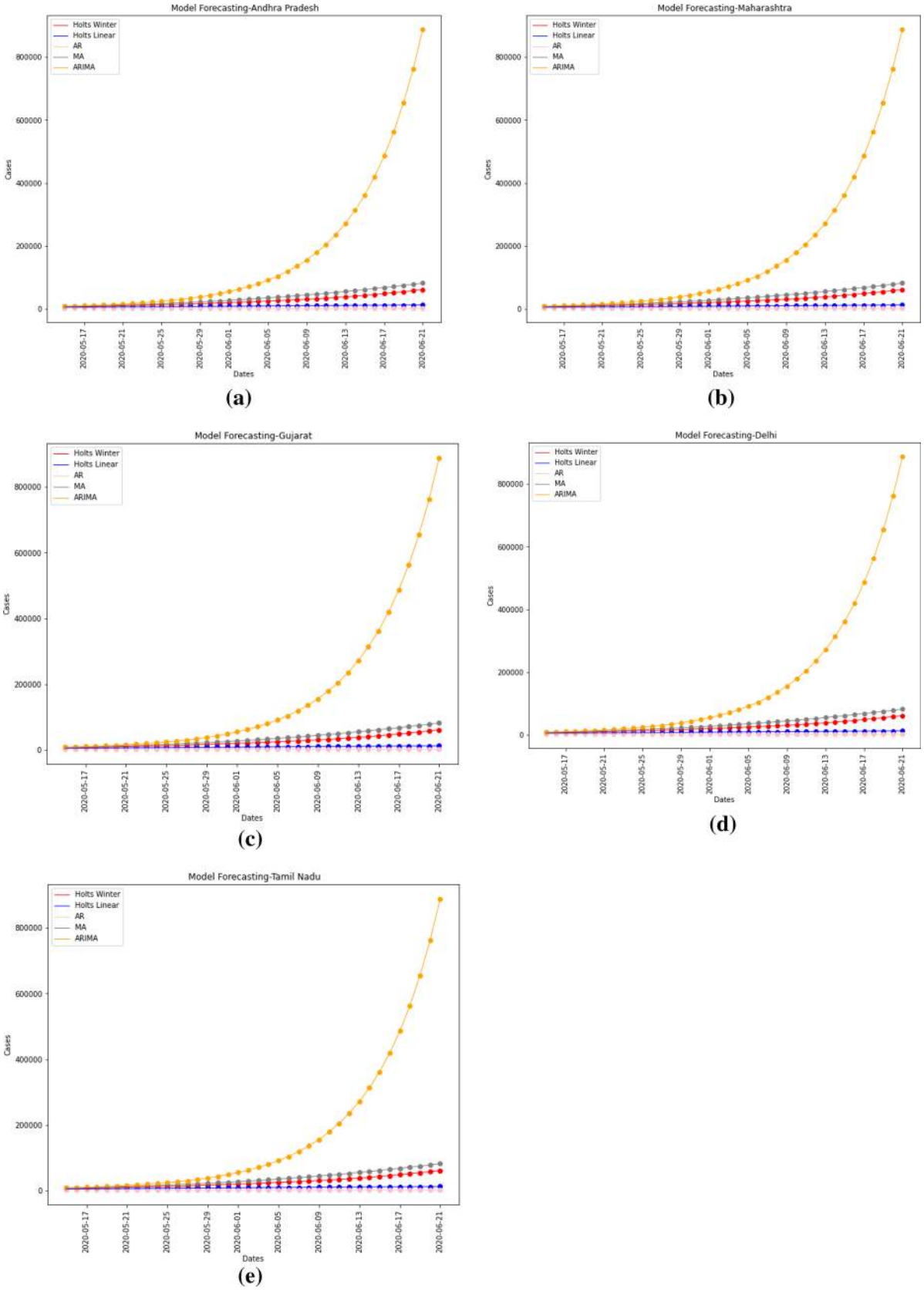

Fig. 4 The outbreak prediction of **a** Andhra Pradesh, **b** Maharashtra, **c** Gujarat, **d** Delhi, and **e** Tamil Nadu

Table 6 Assigned Friedman's rank to the considered models

State	RMSE				
	Holt–Winters	Holt's linear	AR	MA	ARIMA
Andhra Pradesh	76 (1)	162.2 (3)	164.4 (4)	172.3 (5)	147.4 (2)
Maharashtra	338.4 (1)	2919.1 (2)	22,419.6 (5)	12,997.4 (4)	11,184.6 (3)
Gujarat	141.5 (1)	225.2 (2)	11,349.8 (5)	8611.1 (4)	6822.7 (3)
Delhi	425.9 (1)	1264.3 (3)	4931.2 (4)	5291 (5)	486.7 (2)
Tamil Nadu	1991.5 (3)	2554.1 (4)	4027.6 (5)	1491.9 (2)	1378.6 (1)
Avg. rank	1.4	2.8	4.6	4	2.2

Environmental Setup

Here, we discussed the parameter setting of statistical models of Holt–Winters, Holt's Linear, MA, AR and ARIMA model for various states of India like Andhra Pradesh, Maharashtra, Gujarat, Delhi, and Tamil Nadu. The performance of the model is evaluated using the RMSE. Parameter setting in each classifier including bagging is depicted in Table 2.

Results Analysis

The first case in India was reported on January 30, 2020. In the month of February, only three cases were reported and it remained constant throughout the month. From the month of March 2020, the number of cases started increasing steadily. To predict the dynamics of transmission, different time-series statistical models, such as Holt–Winters, Holt's Linear, MA, AR and ARIMA model are simulated on the data that are based on the statistics of India COVID19 [29]. Using above-mentioned models, we forecast the number of confirmed cases for Andhra Pradesh, Maharashtra, Gujarat, Delhi, Tamil Nadu data up to June 21, 2020. The RMSE scores of the Holt–Winters, Holt's linear, MA, AR and ARIMA model are shown in Table 3.

From the Table 3a–d, it can be observed that RMSE, MSE, MAE and MAPE values of Holt–Winters model are less when compared to the RMSE, MSE, MAE and MAPE of other models, such as Holt's Linear, AR, MA and ARIMA model. In Table 3e, it is noted that RMSE, MSE, MAE and MAPE of ARIMA model are less when compared to Holt–Winters model. Further, regarding Holt–Winters model and ARIMA model, Holt–Winters model performed well in four states while the ARIMA model performed well in only one state when compared to Holt–Winters model. Therefore, from the statistics of Andhra Pradesh, Maharashtra, Gujarat, Delhi, it can be stated that RMSE, MSE, MAPE and MAE scores of Holt–Winters statistical model significantly performed well when compared to the scores of Holt's Linear, AR, MA and ARIMA models. Table 4 shows some of the predictions of total number of cases using Holt–Winters, Holt's Linear, AR, MA and ARIMA model with respect to Andhra Pradesh, Maharashtra, Gujarat, Delhi and Tamil Nadu. The predictions are computed up to June 21, 2020. The actual number of cases registered has been depicted in Table 5.

From Tables 4 and 5, it can be noted that the forecast of COVID-19 predicted cases of Holt–Winters model is in proximity with actual values of the registered cases in Andhra Pradesh, Maharashtra, Gujarat, Delhi and Tamil Nadu states. Therefore, it can be concluded that Holt–Winters model performed better predictions of COVID-19 when compared to the other models.

The prediction of number of cases can also be inferred from Fig. 3a–e, which presents the capacity and pattern of each model in the prediction of actual values of COVID-19 cases for Andhra Pradesh, Maharashtra, Gujarat, Delhi, and Tamil Nadu individually. From the Fig. 3a–d, it can be observed that the prediction values of Holt–Winters model, which is represented using red tick, are nearer to the actual validation values of the trained model which are represented using green ticks. In Fig. 3e, the prediction value of ARIMA model, which is represented by yellow ticks, is nearer to the actual validation values of the trained model.

Figure 4a–e represents the prediction of number of confirmed cases by various time series models for Andhra Pradesh, Maharashtra, Gujarat, Delhi, and Tamil Nadu, respectively, from May 15, 2020 to June 21, 2020. From the Fig. 4a–e, it can be inferred that the predictions of number of COVID-19 confirmed cases by the Holt–Winters model

are nearer to the actual value of COVID-19 cases when compared to the other models.

From the above observations, it can be concluded, the Holt–Winters model showed better performance as compared to Holt's Linear, MA, AR and ARIMA model because Holt–Winters model obtains less RMSE, MSE, MAPE, MAE values when compared to other standard models. Moreover, Holt–Winters model performed better than other models and indicated efficient outcomes in terms of RMSE, MSE, MAPE and MAE on the training of 85% data. To statistically validate the results, Friedman test [30] has been performed among the obtained results of all the models over all the considered datasets.

This test considers the average results of all the models in form of ranks (assigned in ascending order as per the performance) [31] and is a non-parametric test. A null hypothesis, "the entire models have similar performance and their differences are merely random", has been considered for conducting this test. Table 6 indicates the assigned ranks (in brackets) to all the models w.r.t. the datasets. By considering all the parameters of Friedman test, "X_F^2" has been evaluated as 16.31. After obtaining X_F^2, the F_F statistic is computed and found to be 10.615. Finally, the critical value is obtained 5.19 which is computed from the F_F statistic and degree of freedom by setting $\alpha = 0.05$ (significance level). The null hypothesis is rejected as the obtained critical value (5.19) is found smaller than the F_F statistic (10.615). Here, the details for process of calculation of the Friedman rank X_F^2, F_F statistic, and critical value can be found [32, 33]. Hence, the proposed model's performance and result are statistically significant and better as compared to other models under the studies. From the test results, it is observed that the performance of the proposed model is statistically significant as compared to other models. By combining the results from the performance metrics, table, and graphs, it is evident that the Holt–Winters method is an efficient model to fit the following growing trend when compared to the other models, such as Holt's Linear, MA, AR and ARIMA models, in forecasting the number of confirmed cases.

Conclusion

Since the first case of COVID-19 in India, the number of registered cases is steadily growing and imposing a great threat to public health in India. In this paper, we employed the Holt–Winters model for forecasting the number of COVID-19 cases in Maharashtra, Tamil Nadu, Gujarat, Delhi, and Andhra Pradesh states of India up to June 21, 2020. The future number of cases has been predicted by analyzing the data from January 30, 2020 to May 14, 2020. The performance of the model has been evaluated using RMSE and the analysis shows that Holt–Winters method has less RMSE, MSE, MAPE and MAE value and generates more accurate predictions when compared with the RMSE, MSE, MAPE and MAE value of Holt's Linear, AR, MA and ARIMA models. From the analysis, it can be predicted that the number of cases in other states of India may also increase in the near future. Based on the predictions, the government has to employ strict policies, such as awareness programs, imposing strict lockdown, etc. to prevent the spread of transmission. Moreover, the government also has to implement necessary measures for enhancing the medical facilities throughout India.

Declarations

Conflict of interest The authors declare that this manuscript has no conflict of interest with any other published source and has not been published previously (partly or in full). No data have been fabricated or manipulated to support our conclusions.

Consent for publication I on behalf of the authors would like to state that the above manuscript is our original research work and it has not been published elsewhere. Also, it has not been submitted to any journal for publication.

References

1. Henderson DA. Smallpox: the death of a disease. Amherst, NY: Prometheus Books; 2009.
2. Spreeuwenberg P, Kroneman M, Paget J. Reassessing the global mortality burden of the 1918 influenza pandemic. Am J Epidemiol. 2018;187(12):2561–7.
3. Yin Y, Wunderink RG. MERS, SARS and other coronaviruses as causes of pneumonia. Respirology. 2018;23(2):130–7.
4. Wuhan Municipal Health and Family Planning Commission. http://wjw.wuhan.gov.cn/front/web/list2nd/no/710
5. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, Cheng Z. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020;395(10223):497–506.
6. Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol. 2019;17(3):181–92.
7. Lai C-C, et al. evere acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): the epidemic and the challenges. Int J Antimicrobial Agents. 2020;55(3):105924.
8. World Health Organization. 2020. Laboratory testing for 2019 novel coronavirus (2019-nCoV) in suspected human cases. Interim guidance, 2 March 2020.
9. Fehr Anthony R, Perlman Stanley. Coronaviruses: an overview of their replication and pathogenesis. Coronaviruses. 2015;1282:1–23.
10. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD. A pneumonia outbreak

associated with a new coronavirus of probable bat origin. Nature. 2020;579(7798):270–3.
11. Marshall M, et al. COVID-19: a danger and an opportunity for the future of general practice. Br J Gen Pract. 2020;70:270–1.
12. Chinazzi M, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science. 2020;368(6489):395–400.
13. Abramowitz M, Stegun IA, editors. Handbook of mathematical functions with formulas, graphs, and mathematical tables, vol. 55. Wasington: US Government Printing Office; 1964.
14. Verma B, Ray SK, Srivastava RN. Mathematical models and their applications in medicine and health. Health Popul Perspect Issues. 1981;4(1):42–58.
15. Adomian G. Solving the mathematical models of neurosciences and medicine. Math Comput Simul. 1995;40(1–2):107–14.
16. Castillo-Chavez C, Castillo-Garsow CW, Yakubu AA. Mathematical models of isolation and quarantine. JAMA. 2003;290(21):2876–7.
17. Massad E, Wilder-Smith A. Risk estimates of dengue in travelers to dengue endemic areas using mathematical models. J Travel Med. 2009;16(3):191–3.
18. Earnest A, Tan SB, Wilder-Smith A, Machin D. Comparing statistical models to predict dengue fever notifications. Comput Math Methods Med. 2012. https://doi.org/10.1155/2012/758674.
19. Yoo C, Ramirez L, Liuzzi J. Big data analysis using modern statistical and machine learning methods in medicine. Int Neurourol J. 2014;18(2):50.
20. Getoor L, Rhee JT, Koller D, Small P. Understanding tuberculosis epidemiology using structured statistical models. Artif Intell Med. 2004;30(3):233–56.
21. Xu H, Yan L, Qiu CM, Jiao B, Chen Y, Tan X, Chen Z, Ai L, Xiao Y, Luo A, Li S. Analysis and prediction of false negative results for SARS-CoV-2 detection with pharyngeal swab specimen in COVID-19 patients: a retrospective study. MedRxiv. 2020. https://doi.org/10.1101/2020.03.26.20043042.
22. Kucharski AJ, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. Lancet Inf Dis. 2020;20(5):553–8.
23. Gao Y, Zhang Z, Yao W, Ying Q, Long C, Fu X. Forecasting the cumulative number of COVID-19 deaths in China: a Boltzmann function-based modeling study. Infect Control Hosp Epidemiol. 2020;2:1–3.
24. Roques L, Klein EK, Papaix J, Sar A, Soubeyrand S. Using early data to estimate the actual infection fatality ratio from COVID-19 in France. Biology. 2020;9(5):97.
25. Poonia N, Azad S. Short-term forecasts of COVID-19 spread across Indian states until 1 May 2020. arXiv preprint. arXiv:2004.13538. 2020.
26. Zhang GP. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing. 2003;50:159–75.
27. Singh RK, Rani M, Bhagavathula AS, Sah R, Rodriguez-Morales AJ, Kalita H, Nanda C, Sharma S, Sharma YD, Rabaan AA, Rahmani J. Prediction of the COVID-19 pandemic for the top 15 affected countries: Advanced autoregressive integrated moving average (ARIMA) model. JMIR Public Health Surv. 2020;6(2):e19115.
28. Koehler AB, Snyder RD, Ord JK. Forecasting models and prediction intervals for the multiplicative Holt-Winters method. Int J Forecast. 2001;17(2):269–86.
29. Coronavirus Outbreak in India—Covid19india.Org. https://www.covid19india.org/
30. Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc. 1937;32:675–701.
31. Friedman MA. Comparison of alternative tests of significance for the problem of m rankings. Ann Math Stat. 1940;11:86–92.
32. Naik B, Nayak J, Behera HS, Abraham A. A self adaptive harmony search based functional link higher order ANN for nonlinear data classification. Neurocomputing. 2016;179:69–87.
33. Nayak J, Naik B, Behera HS, Abraham A. Elitist teaching-learning-based optimization (ETLBO) with higher-order Jordan Pi-sigma neural network: a comparative performance analysis. Neural Comput Appl. 2018;30(5):1445–68.
34. Chen X, Yu B. First two months of the 2019 Coronavirus Disease (COVID-19) epidemic in China: real-time surveillance and evaluation with a second derivative model. Global Health Res Policy. 2020;5(1):1–9.
35. Gupta R, Pandey G, Chaudhary P, Pal SK. SEIR and regression model based COVID-19 outbreak predictions in India. medRxiv. 2020. https://doi.org/10.1101/2020.04.01.20049825
36. Al-qaness MA, Ewees AA, Fan H, Abd El Aziz M. Optimization method for forecasting confirmed cases of COVID-19 in China. J Clin Med. 2020;9(3):674.
37. Tandon H, et al. Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future. arXiv preprint. arXiv:2004.07859 (2020).
38. Chintalapudi N, Battineni G, Amenta F. COVID-19 virus outbreak forecasting of registered and recovered cases after sixty day lockdown in Italy: a data driven model approach. J Microbiol Immunol Infect. 2020;53(3):396–403.
39. Kumar P, Singh RK, Nanda C, Kalita H, Patairiya S, Sharma YD, Rani M, Bhagavathula AS. Forecasting COVID-19 impact in India using pandemic waves nonlinear growth models. medRxiv. 2020. https://doi.org/10.1101/2020.03.30.20047803
40. Anne R. ARIMA modelling of predicting COVID-19 infections. medRxiv. 2020. https://doi.org/10.1101/2020.04.18.20070631