Int J Syst Assur Eng Manag (December 2022) 13(6):2920–2933
https://doi.org/10.1007/s13198-022-01762-7

ORIGINAL ARTICLE

Time series forecasting and mathematical modeling of COVID-19 pandemic in India: a developing country struggling to cope up

Vidhi Vig[1] · **Anmol Kaur**[2]

Received: 4 September 2021 / Revised: 5 August 2022 / Accepted: 6 August 2022 / Published online: 23 August 2022
© The Author(s) under exclusive licence to The Society for Reliability Engineering, Quality and Operations Management (SREQOM), India and The Division of Operation and Maintenance, Lulea University of Technology, Sweden 2022

Abstract COVID-19 has spread around the world since it begun in December 2019. The pandemic has created an unprecedented global health emergency since World War II. This paper studies the impact of pandemic and predicts the anticipated casualty rise in India. The data has been extracted from the API provided by https://www.covid19india.org/ and covers up the time period from 30th January 2020 when the first case occurred in India till 13th January 2021. The paper provides a comparative study of six machine learning algorithms namely SMOreg, Random Forest, lBk, Gaussian Process, Linear Regression, and Autoregressive Integrated Moving Average (ARIMA) in forecasting deceased COVID 19 cases, via the data mining tool such as Weka and R. The major findings show that the best predictor model for anticipating the frequency of deceased cases in India is ARIMA (5,2,0). Utilizing this model, we estimated the propagation rate of deceased cases for the next month. The findings reveal that the fatal cases in India could rise from 151,174 to 157,179 within one month with an average of 190 death reports every day. This study will be helpful for the Indian Government and Medical Practitioners in assessing the spread of pandemic in India and devising a combat plan to mitigate the pandemic.

Keywords COVID-19 · Time series · Forecasting · ARIMA · Machine learning · Classifiers

✉ Anmol Kaur
anmolkaur1037@gmail.com

[1] Department of Computer Science, University of Delhi, Delhi, India

[2] Shri Guru Tegh Bahadur Khalsa College, University of Delhi, University Enclave, New Delhi, Delhi 110007, India

Abbreviations

AR	Autoregressive
MA	Moving average
ARIMA	Autoregressive integrated moving average
COVID-19	Coronavirus disease 2019
ADF	Augmented Dickey–Fuller
RMSE	Root mean square error
MAE	Mean absolute error
MAPE	Mean absolute percentage error
RMSRE	Root mean squared relative error
GPC	Gaussian process classifier
AICc	Corrected Akaike's information criterion
WEKA	Waikato environment for knowledge analysis

1 Introduction

Towards the end of 2019, a new disease of unknown etiology appeared in China. It was later distinguished to be as a modern strain of coronavirus (nCoV), hence named the "COVID-19 virus". Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), generally acknowledged as COVID-19, is a novel coronavirus, which was first identified amid an outbreak of respiratory disease in Wuhan city in Hubei Province, China. Coronaviruses belong to a family of viruses that were first discovered in the mid-1960s. Seven recognized types of coronaviruses can infect humans. Of the seven, four can provoke mild respiratory infections, and the other two types, Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) and the Middle East Respiratory Syndrome Coronavirus (MERS-CoV), cause critical respiratory infections. The seventh coronavirus type, SARS-CoV-2 is responsible for the ongoing global pandemic. SARS-CoV-2 begins with symptoms of high temperature and a tenacious dry cough (Coronaviruses 2020). Since the incubation phase

of the novel coronavirus is around 2–14 days after exposure, a large fraction of people stays asymptomatic and oblivious.

Governments all around the world implemented several prudent measures to constraint the spread of the virus. The last year saw multiple restraints including nation-wide lockdowns, closing of air-travel between nations, prohibition of public events, etc. However, the disease struck the world off guard, and spread across the globe rapidly within a span of weeks. On March 11, 2020 COVID-19 was pronounced as a pandemic which by then had surpassed 118,000 cases in 114 countries (World Health Organization 2020). This disease has dramatically spread over the world and had a profound impact on healthcare system in many countries. Over a year after the first case more than 92 million cases have been confirmed, with more than 1.97 million deaths attributed to COVID-19, across 219 countries and territories across the world (as of 13 January 2021).

India, as well as other countries, has been hit very hard by the Covid-19 virus. India witnessed the first case of COVID-19 at the end of January 2020 in Kerala's Thrissur district. Three more cases were reported in February, all were students returning from Wuhan. In March, the transmissions grew after several people with a tour history to affected countries tested positive. On 12 March 2020, a 76-year-elderly person turned into India's first COVID-19 casualty. As an approach to forestall the spread of COVID-19 in India, the government imposed a 21-day nationwide lockdown starting 23 March 2020, which was later extended. During the lockdown, the whole country was divided into different containment zones (red, orange, and green). This helped the officials to adequately manage the resources, and address the public health & containment process during the pandemic. The districts where the risk of transmission was high and had multiplying rates were marked as red zone, districts that reported no confirmed case in the last 21 days were marked as green zone and areas that are included in neither of the zones as the orange zone. The lockdown prolonged to May 31 and some economic activities were permitted in cities with low transmission rate. This process of mitigation was also expanded, enabling provinces and regional authorities to directly control community guidelines in the light of developments in their area of jurisdiction.

Figures 1 and 2 show the number of daily and total deceased cases reported in India from 3 March, 2020 to 13 January, 2021, respectively. India's daily casualty records followed a bell-shaped trajectory. The months August, September, and October reported more than 1,000 deaths daily. In January 2021, this number dropped to less than 300. Due to significant delay in reporting of deaths, often estimates from several days were reported altogether. This resulted in some sudden spikes in the graph.

The SARS-COV-2 outbreak has deeply impacted the economy, environment, social and health structure of the

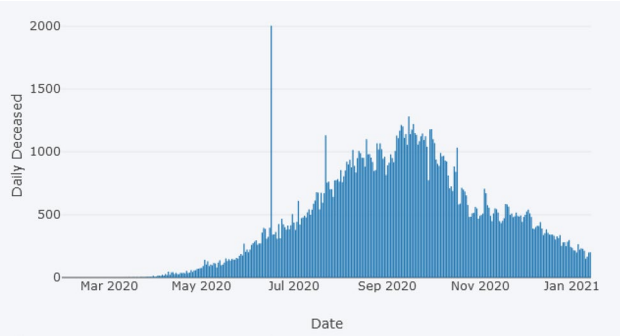

Fig. 1 Daily deceased cases of COVID-19 reported in India till January, 2021

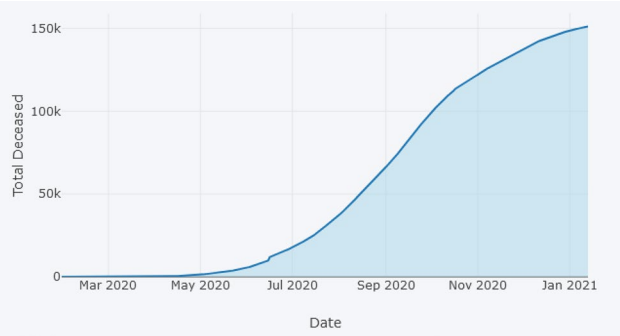

Fig. 2 Total deceased cases of COVID-19 reported in India till January, 2021

world. Due to the pandemic human activities such as travel decreased and consecutively the world experienced a drop in air pollution and water pollution. Srivastava (2021) found that air quality improved during lockdown with a critical decrease in concentration of common pollutants. Ghosh and Ghosh (2020) evaluated 15 empirical research articles from four continents and found that concentration level of CO, NO, NO_2, NO_X, NH_3, SO_2, PM_{10}, $PM_{2.5}$ decreased and that of O_3 increased during lockdown. Gupta et al. (2020) analyzed the data of various harmful pollutants in the environment along with satellite images from National Aeronautics and Space Administration for comparison of different parameters. Mahato et al. (2020) observed a decrease in temperature by 15-degree Celsius in India. Delhi, which is deemed to be one of the most polluted cities in India observed significant reduction in National Air Quality Index (NAQI) during lockdown.

Even though the lockdown had some positive effect on our environment, the pandemic caused economic and social disruption. The lockdown gave rise to shut businesses, unemployment and crumbling incomes. In India, the vulnerable population fell deeper into the abyss of poverty because of extreme lockdown measures. People who relied

on informal job sector lost their jobs, leaving them and their families on the verge of starvation. The stress and panic resulting from loss of a job or uncertain future and financial turmoil leads to a major setback in the mental health of the economically vulnerable.

India's inconsistent health facilities, flickering economic and social disparities had made lockdown a hard course for the vulnerable components of the society. The nationwide lockdown caused major economic loss and crippled the country's large population of breadwinners and migrant laborers and lead to a setback in the mental health front. History suggests that disease pandemics have led to drastic psychological repercussion in the long run. Over a year after the Ebola epidemic in the year 2014, nearly 50% of all respondents exhibited at least one sign of anxiety or depression (Jalloh et al. 2018). The study found that people with any degree of exposure to Ebola were likely to be vulnerable to anxiety-depression and Post-Traumatic Stress Disorder (PTSD). A similar pattern can be seen in the global HIV pandemic. Studies have found a high seroprevalence of chronic mental illnesses in HIV patients (World Health Organization 2008). Since the health system grapples to save millions of people daily, there is a probable risk of deteriorating mental health due to the pandemic (Roy et al. 2021; Kumar and Nayar 2021). To wipe out any repercussion due the COVID-19 pandemic, the mental health system needs to be given equal emphasis along with other schemes to administer the pandemic in general.

Vaccines are proved to be the most fruitful and cost-effective means to avert and control contagious diseases. On January 9, 2021, the Govt. of India announced the Covid-19 vaccination drive. The national campaign would begin on January 16, healthcare workers and the frontline workers (approximately 30 million people) will be given top priority, followed by those over the age of 50 and the under-50 groups with co-morbidities numbering around 270 million.

Vaccinations are the secure way to implement herd immunity in a population. To reduce the overall spread of the virus, a significant number of people will need to be vaccinated. One of the objectives to achieve herd immunity is to keep vulnerable groups who are unable to get vaccinated, safe and protected. The safety records of the vaccines are promising so far, but the coming months will present a broad view as the sample size increases. Even when herd immunity is accomplished, continuous monitoring, revaccination, and treatment of isolated cases will still be required to restrain COVID-19.

In such a case, modeling the current situation and determining the outcome of the future is crucial to utilize our resources wisely. This study aims to scrutinize the trend of COVID-19 cases in India and foresee its evolution. Since the disease is dynamic, no model can guarantee absolute validity for the future. Therefore, we'd like to develop an understanding of the present state of the pandemic. This paper presents a comparative study between six machine learning algorithms namely SMOreg, Random Forest, lBk, Gaussian Process, Linear Regression, and Autoregressive Integrated Moving Average (ARIMA) in forecasting deceased COVID 19 cases. The algorithms have been assessed on the COVID-19 dataset provided by https://www.covid19india.org/ recorded from November 1, 2021, till January 13, 2021. This study will help us to grasp the spread of SARS-CoV-2 in India better and help the government and the public to ideally utilize the assets accessible to them.

2 Related work

Time-series forecasting helps us to predict future events based on past data using machine learning algorithms. In decisions entailing risks, time series is perceived as one of the productive ways of making predictions. This is well documented in the literature proposed by Mahalakshmi et al. (2016) and De Gooijer et al. (2006). Amid the many time series models, the ARIMA models are particularly attractive. The application of time series models is manifold including agricultural productivity, stock price returns, index predictions, etc. Vijayarani et al. (2021) used ARIMA model to analyze the top five crime events namely theft, assault, criminal damage, battery, and fraud in Chicago city. This study aims to find the top crimes on a monthly, weekly, and daily basis and analyze the past data to reduce the future crime count in Chicago. They compared an existing ARIMA model and the proposed enhanced ARIMA model using Mean Absolute Error (MAE), Mean Square Error (MSE), and Root of Mean Square Error (RMSE) as performance metrics. Dimri et al. (2020) used seasonal ARIMA model to analyze climatic changes and the precipitation for the Bhagirathi river situated in Uttarakhand, India. Using 28 years (from January 1987 to December 2015) of monthly pollutants data.

Barman (2020) used LSTM and ARIMA algorithms to build the prediction model for COVID-19 confirmed cases in four countries—United States, Italy, Spain, and Germany. They applied multiple LSTM architectures like vanilla LSTM, CNN LSTM, and Bidirectional LSTM. The authors also proposed k-period performance metrics to gauge the performance of time series models. They concluded that LSTM models perform comparably with the ARIMA model and pointed out that each model has its benefits and limitations.

Benvenuto et al. (2020) employed the ARIMA model to foresee the incidence and prevalence of Covid-19 using the John Hopkins epidemiological data. The authors have used the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots to choose the best parameters for

the model. ARIMA(1,0,4) and ARIMA(1,0,3) models were selected as the best for the prevalence and incidence, respectively. Istaiteh et al. (2020) applied ARIMA, ANN, LSTM, and CNN models to COVID-19 data from January 22, 2020, to June 30, 2020, to forecast the confirmed Covid-19 cases for 189 countries around the world for the next seven days. The authors have used MAPE, RMSLE, and MSLE error measurements to compare the four models.

Pandey et al. (2020) employed SEIR and regression models to investigate the spread of COVID-19 in India from January 30, 2020, to March 30, 2020. The accuracy of the models was determined using Root Mean Squared Log Error (RMSLE). Alzahrani et al. (2020) estimated the rise in COVID-19 cases before the 2020 Hajj in Saudi Arabia. They compared AR, MA, ARMA, and ARIMA models using RMSE, MAE, MAPE, R^2, and RMRSE to ascertain the best model. The findings showed that ARIMA(2,1,1) was the best model fit (with RMSE of 21.17 and MAE of 14.93) to forecast the expected rise in COVID-19 cases in Saudi Arabia. The study found that the cases will rise with a growth rate varying from 0.2 to 10.8 daily in Saudi Arabia. Ceylan (2020) developed ARIMA models to estimate the pattern of COVID-19 in Italy, Spain, and France from February 2020 to April 2020. The models ARIMA(0,2,1), ARIMA(1,2,0) and ARIMA(0,2,1) were chosen based on MAPE to foresee the pervasiveness of the pandemic in Italy, Spain, and France. Gupta and Pal (2020) applied the ARIMA(1,1,2) model and Exponential Smoothing to John Hopkins University COVID-19 data repository from January 30, 2020, to March 24, 2020, and estimated the trend of the pandemic in India in the upcoming month.

Using the historical demand data from January 2010 until December 2015, Fattah et al. (2018) developed several ARIMA models, and the best fit was chosen according to four measures: Akaike criterion, maximum likelihood, Bayesian criterion, and standard error. Although ARIMA models are quite adaptable, often hybrid ARIMA models are proposed to exploit the capability of the conventional time series model. To enhance the precision of the ARIMA model for annual runoff time series forecasting, Wang et al. (2015) suggested the use of ARIMA model coupled with ensemble empirical mode decomposition (EEMD). Faruk (2010) presents ARIMA and ANN models to predict the water quality at Büyük Menderes River, Turkey.

Using the COVID-19 Symptoms and Presence dataset from Kaggle, Villavicencio et al. (2021) developed a model to predict the existence of COVID-19 in a person. WEKA software was used to implement the J48 Decision Tree, Nave Bayes, Support Vector Machine, Random Forest and K-Nearest Neighbors algorithms. The efficacy of each model was reviewed using tenfold cross validation, kappa, mean absolute error and other major accuracy measures. The findings demonstrate that Pearson VII universal kernel-based Support Vector Machine outperform other methods by achieving 98.81% accuracy and MAE of 0.012. To predict the number of cases in four nations, Abolmaali and Shirzaei (2021) employed logistic function, linear regression, Susceptible-Infected-Recovered (SIR) model, and ARIMA model. The models are evaluated using MSE. The ARIMA model excelled the other three models since it was able to present the prediction with the least amount of error. Alabdulrazzaq et al. (2021) uses Kuwait as a case study to determine the accuracy of ARIMA model in predicting the infected and recovered COVID-19 cases. A wide variety of accuracy measures such as RMSE, MAE, AIC, MAPE and so on were used to evaluate the model. Their ARIMA model's accuracy in predicting values was in sync with what was actually seen over the specified interval.

Some researchers have looked for signs that the COVID-19 data exhibit seasonality. ArunKumar et al. (2021) investigated the pandemic trend for the top sixteen nations, which accounted for 70–80% of global infections. The data were analysed using the time series models ARIMA and Seasonal ARIMA (SARIMA), and the evaluation metrics MAE, MSE, RMSE, and MAPE were utilised to compare the results. The SARIMA model's predictions are consistent with the seasonality found in the COVID-19 data. Another study by Konarasinghe (2021) trained the Sama Circular Model (SCM) and SARIMA to investigate the outbreak in Italy. SARIMA proved less effective than SCM at capturing the seasonal behaviour.

In this study we have used SMOreg, Random Forest, IBk, Gaussian Process, Linear Regression, and Autoregressive Integrated Moving Average (ARIMA) to forecast the deceased COVID 19 cases in India. The data used is scraped from https://www.covid19india.org/ and the performance of models is evaluated using RMSE, MAPE and MAE. Figure 3 shows the conceptual framework of the proposed study.

3 Materials and methods

Time series analysis (Wei 2013) and dynamic modeling is an exciting research field with a multitude of applications in several areas including business, economics, finance, and computer science. It studies the past estimates of time series and defines a model to foresee future events. The time series estimation results have a crucial role in many branches of applied sciences, so it is essential to emphasize on refining the accuracy of a model. Time series forecasting is customarily performed using ARIMA models. The ARIMA model is established on the Box-Jenkins algorithm (Box et al. 2015) which is widely used in many areas of research to predict future events.

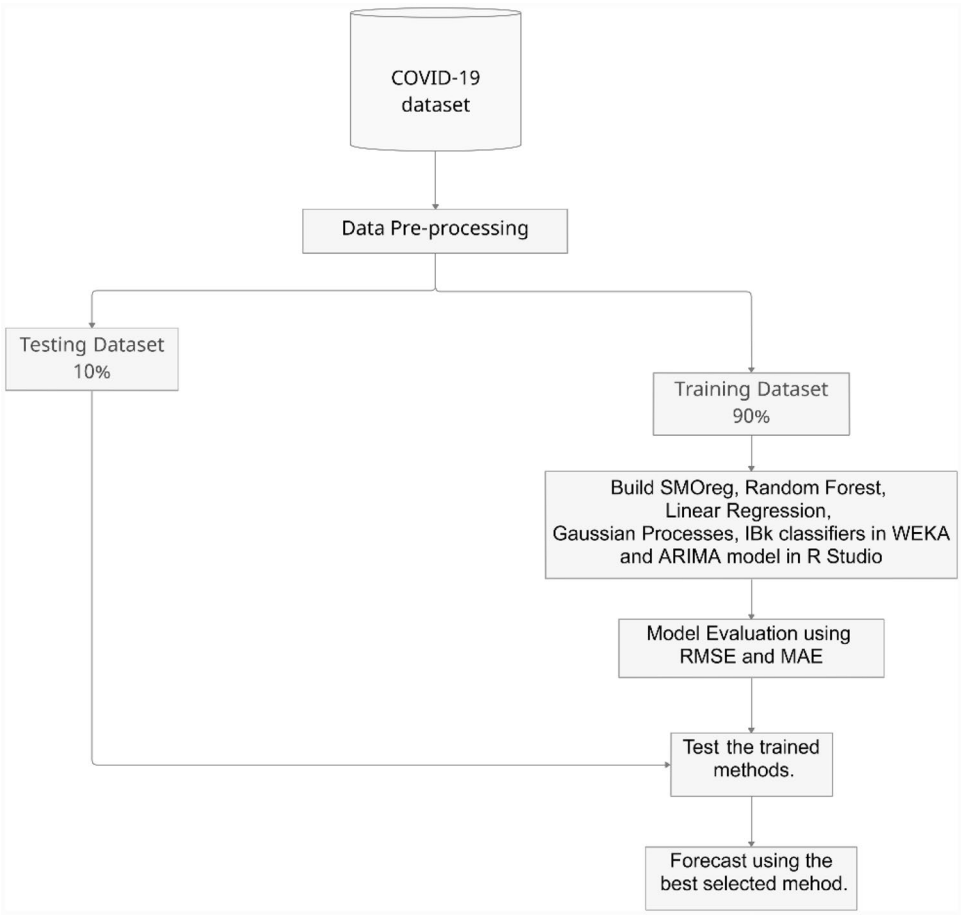

Fig. 3 Conceptual framework of the proposed forecasting methods

3.1 ARIMA model

Autoregressive Integrated Moving Average Model (ARIMA) is a comprehensive model of the Autoregressive Moving Average (ARMA) that joins the Autoregressive (AR) process and Moving Average (MA) process and assembles a consolidated model of the time series. ARIMA(p, d, q) catches the fundamentals of the model:

- AR (Autoregression): It shows the dependency between an observation and past observations (p).
- I (Integrated): Substitutes data values with differenced data of 'd' order to make the time series stationary.
- MA (Moving Average): It considers the reliance among observations and the residual error terms (q).

Effectively combining the Autoregressive (AR) and Moving Average (MA) models forms a simple and useful class of time series models called ARMA models. ARIMA can deal with non-stationary time series data because of its "integrate" step. The "integrate" part comprises of differencing the time series to convert a non-stationary time series to a stationary time series.

3.2 Classification

Classification is the method of recognizing, distinguishing, and understanding ideas and objects. The term "classifier" also attributes to the mathematical function, implemented by a classification algorithm that categorizes input data. In machine learning, classification is viewed as an example of supervised learning, i.e., learning where a set of correctly identified observations is available. In this study, we analyze five classifiers namely Linear Regression, Gaussian Processes, IBk, Random Forest, and SMOreg.

3.2.1 Random forest

Random Forest was developed by Leo Breiman. It classifies based on the results attained from the decision trees it produces while training. Trees have low bias and high variance so they tend to overfit data but Random Forests generate decision trees on random samples of the training data which reduces the variance, thereby improving its performance. It is easy to develop, easy to implement, and generates robust classification. It can handle missing values, continuous, categorical and binary data which makes it apt

for high dimensional data modeling. Random Forest tends to outperform most classifiers in respect of accuracy without any issue of overfitting. The major advantages of Random Forests are their high accuracy, potential to determine variable importance and non-parametric nature.

3.2.2 Linear regression

This classification method utilizes linear regression to measure the distance from a sample to a specific class. Linear regression presumes linear relationship between the input and output variable. It uses the Akaike criterion for model selection, and is able to deal with weighted instances.

3.2.3 lBk

lBk classifiers are a simple and effective approach that belongs to a lazy learning category. They are also known as k nearest-neighbors (k-NN). It does not learn immediately from the training set instead it stores the dataset and during classification, it performs an action on the dataset. In such a technique, the model does not require learning and the prediction can be obtained from the raw training instances. lBk uses a majority vote between the new instance and the k most similar instances, where the distance is the main factor in determining the similarity between two data points. This moderates down significantly as the volume of information increments, making it an illogical choice in circumstances where forecasts have to be be made faster.

3.2.4 Guassian processes

The Gaussian process classifier (GPC) is a kernel-based classifier. It is a generalization of Gaussian Process (GP). GP is a statistical model that has a Bayesian approach to classification and regression problems. GPC has three major advantages. The first advantage is that it can handle high-dimensional and nonlinear issues. Second advantage is GPC also offers probabilistic outputs instead of determinant classification results. The third advantage is that being a non-parameterized model, GPC can also use evidence to implement a model selection process in a fully automatic manner.

3.2.5 SMOreg

The SMOreg represents Sequential Minimal Optimization Regression. It is an implementation of Support Vector Machine (SVM) for regression. SVM is trained using SMOreg. The training of SVM requires the answer for exceptionally enormous quadratic programming optimization problem. SMO first partitions the enormous quadratic programming problem into series of quadratic programming problem units which are then addressed analytically.

3.3 Parameters used to measure performance of model

Performance metrics assess the accuracy and choose the best forecasting method. The performance metrics used in this paper are as follows:

3.3.1 Mean absolute error (MAE)

MAE is used to determine how close the forecasted values and eventual outcomes are.

$$\mathrm{MAE} = \Sigma \mathrm{i}_{=1}^{\mathrm{N}} (z_i - \hat{z}_i)/N \tag{1}$$

where N is the number of data points, x_i is the actual observation and xÎ is the predicted value.

3.3.2 Root mean square error (RMSE)

RMSE is defined as standard deviation of the differences between values predicted and the values observed.

$$\mathrm{RMSE} = \sqrt{(\Sigma \mathrm{i}_{=1}^{\mathrm{N}} (z_i - \hat{z}_i)^2 / N)} \tag{2}$$

3.3.3 Mean absolute percentage error (MAPE)

MAPE is another measure of prediction accuracy of a forecasting model. It is defined as:

$$\mathrm{MAPE} = (1/N)\Sigma \mathrm{i}_{=1}^{\mathrm{N}} (\hat{z}_{i=} z_i)/\hat{z}_i \tag{3}$$

The terms in Eqs. (2) and (3) have the same meaning as in Eq. (1).

3.4 Dataset

In this paper, the data from https://www.covid19india.org/ is incorporated. The data of COVID-19 cases in India is available through their Github API in both CSV and JSON formats. We downloaded the 'case_time_series' CSV file for convenience. The COVID-19 confirmed, recovered, and deceased cases in India from January 30, 2020 to January 13, 2021 are documented daily and cumulatively in the dataset. India's winter season lasts from November to February. The dataset is then preprocessed in R and truncated to observations from November 2021 to January 2021 in order to forecast the deceased cases for the future month and to reduce the impact of seasonal variations on the data.

We used R (Team 2013) for constructing the ARIMA model. The data were cleaned and processed using R library called tidyverse (Wickham et al. 2019). We analyzed the COVID-19 data and performed data visualization

using plotly, which gave a complete idea of the brief summarization of our dataset. We have used this package for plotting our dataset results. For time series forecasting, we used R modules like forecast (Hyndman et al. 2020) and tseries (Trapletti et al. 2020).

Waikato Environment for Knowledge Analysis (Weka) (Weka 2011), an open-source software tool is used for implementation of machine learning algorithms. In this paper, we use SMOreg, Random Forest, IBk, Gaussian Processes and Linear Regression classifiers implemented in Weka. Weka's Explorer GUI and RStudio are used for code development and testing.

3.5 Data preprocessing

Data preprocessing is time-consuming, but it is also a crucial step in machine learning. Realistic data typically comes from various platforms and can be noisy, obsolete, incomplete, and inconsistent. Therefore, it is essential to recast raw data into an appropriate format for analysis and insight. In this paper, the preprocessing step comprises eliminating outliers and redundant instances. Output from the data preprocessing step is the final training set. Key data cleansing steps include consideration of missing data and duplicate observations, which may occur during data collection at the source. These must be dealt with judiciously.

The dataset has been prepared and cleaned where only appropriate attributes were obtained from the original dataset. The extracted dataset has only 74 observations between November 1, 2020 and January 13, 2021. We have analyzed and forecasted the total deceased time series.

Managing outliers across the data set is another tricky problem to solve. An outlier is an extreme value that lies in a data series. This value can be very small or quite large and thus can affect the overall observation made in the data series. We have used Whisker's plot to explicitly determine outliers in our dataset. The Whisker's plot, often referred as the box plot, is a graphical method helpful in determining outliers. It defines the upper and lower limit for a dataset. If a data value falls beyond these boundaries, it will be regarded as an outlier. Using this method, we can easily identify outliers and handle them judiciously before making any further observations. This helps generate more precise results which are not affected by any extremes or unusual values. The boxplot in Fig. 4 shows that the total deceased cases range from 122,050 to 151,174 with 131,609 as lower quartile, 140,606 as median, 147,068 as upper quartile, and no outlier.

The next step in data preprocessing is to test if the data is normally distributed. The data distribution determines the set of statistical tests that can be applied. Shapiro–Wilk Test is one way to check normality (Shapiro and Wilk 1965). In this test, the null hypothesis being tested is that a set of N

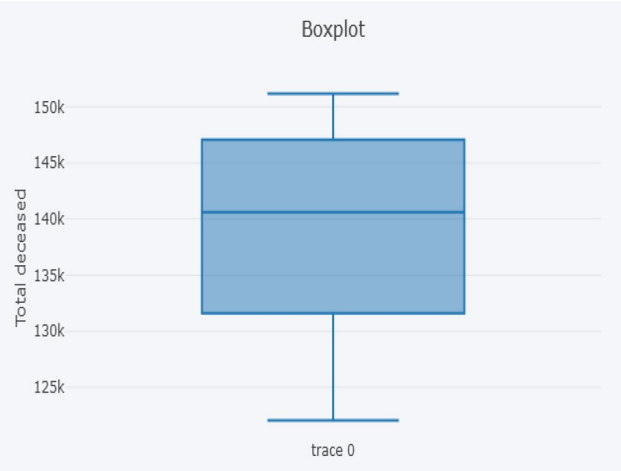

Fig. 4 Boxplot of total deceased cases in India

observations is normally distributed. Hypothesis tests use a p-value to confirm or discard the null hypothesis. If the p-value is small, it strengthens the evidence to discard the null hypothesis. Here, p-value of the Shapiro–Wilk test for total deceased data is 0.08475, i.e., p-value > 0.05 which supports the null hypothesis; therefore, the data is normally distributed.

Table 1 presents the basic statistics of the pre-processed data.

3.6 Time series model fitting

The primary purpose of using time-series models, such as ARMA models, is to predict. ARIMA modeling of time-series data utilizes three steps: model identification, parameter estimation, and diagnostic testing. To identify the candidate models for the best fit, differencing of the time series is done to attain stationarity. The Augmented Dickey-Fuller (ADF) unit-root test (Dickey and Fuller 1981) is commonly used to check whether the time-series is stationary or not. If a time-series becomes stationary after differencing, then the ARMA model becomes ARIMA model where "I (integrated)" is the order of differencing. Dickey-Fuller statistic is 2.4772 with a p-value of 0.99 for the total deceased time series which indicates non-stationarity. From Fig. 5, we conclude that the ideal value of d is 2 for total deceased time series. Subsequent parameters p (order of AR model) and q (order of MA model) are determined using ACF and PACF functions plots (Fig. 6). The autocorrelation function maps the correlation between the actual data values with its lagged values while the partial autocorrelation maps the correlation between a present value with the lagged value.

We then determine the optimal ARIMA parameters (p,d,q) using the AICc (corrected Akaike's Information Criterion). It provides a relative measure of different models, by

Table 1 Descriptive statistics of the data

	Daily confirmed	Daily recovered	Daily deceased	Total confirmed	Total recovered	Total deceased
Min	12,481	16,737	150	8,229,245	7,542,738	122,050
Max	50,465	58,524	707	10,512,755	10,146,102	151,174
Median	30,518	37,700	412	9,719,900	9,196,442	140,606
Mean	31,479	35,901	400.2	9,583,653	9,069,938	139,111
Standard deviation	11,176.728	11,466.863	130.048	684,713.77	776,804.356	8831.73
Variance	124,919,254.5	131,488,961.9	19,509.130	468,832,956,108.36	603,425,008,284.8	77,999,529.5

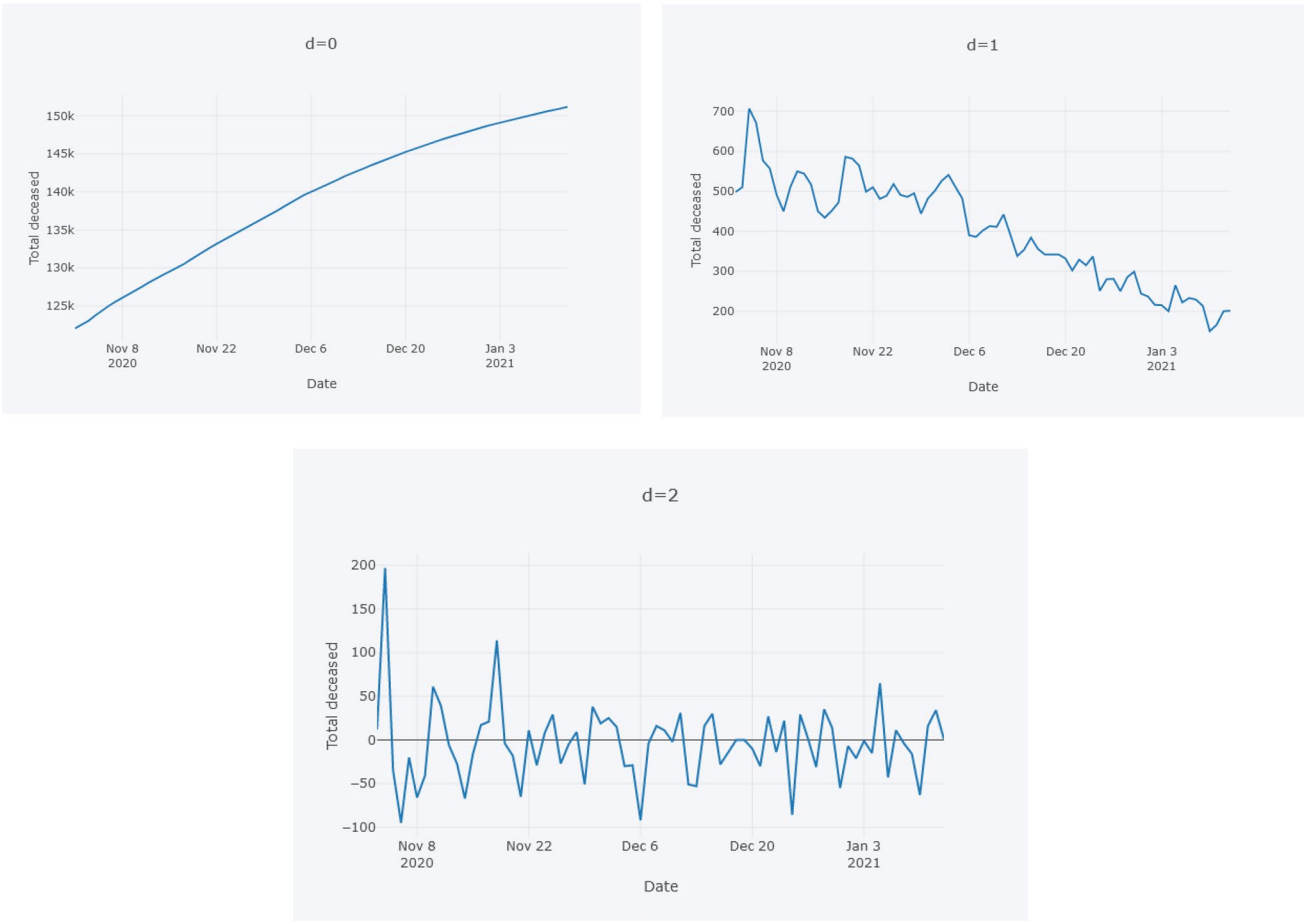

Fig. 5 Difference plots of deceased COVID-19 cases

comparing how well they fit the data. We consider potential combinations of (p,d,q) and evaluate the AICc (Table 2). The model with minimal AICc value selected as the best-fitted ARIMA model. ARIMA (5, 2, 0) is the best-fitted model for the COVID-19 total deceased cases in India (see Table 3). ARIMA (5, 2, 0) parameters, MAE, RMSE and MAPE are shown in Table 3.

The efficacy of the model is validated with the help of the Ljung-Box test (Ljung et al. 1978). The Ljung-Box test examines the lack of fit of a time series model. It inquires the null hypothesis that autocorrelations up to lag k are zero. If the p-value is greater than the critical value, auto-correlations for one or more lags may differ significantly from zero, indicating that values are not arbitrary and independent over time (i.e., the model is inadequate). The p-value is 0.1032 for the time series which is greater than the specified critical value (0.05), therefore the selected ARIMA model does not show a lack of fit. The scatter plot, histogram, and auto–correlation function (ACF) plots of the residuals presented in Fig. 7 show how well the

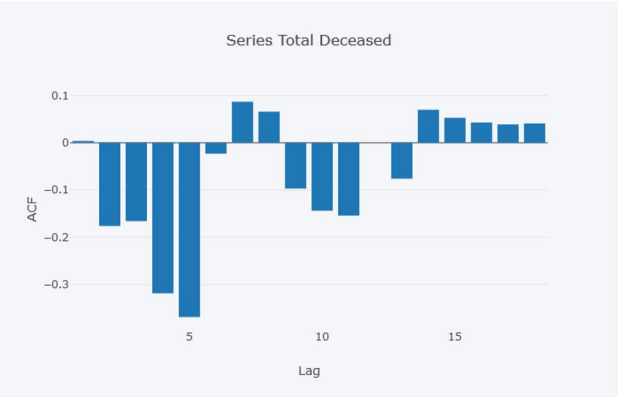

Fig. 6 ACF and PACF plots for deceased COVID-19 cases

Table 2 ARIMA model selection based on AICc values

Model			AICc value
p	d	q	
1	2	0	753.9145
1	2	1	752.0297
1	2	2	750.9312
1	2	3	752.9648
2	2	0	753.5876
2	2	1	749.5747
2	2	2	740.5094
2	2	3	754.6795
3	2	0	753.9174
3	2	1	750.1587
4	2	0	748.1709
4	2	1	746.5024
5	2	0	740.1521

Table 3 Estimates of ARIMA (5,2,0) parameters and its performance

Coefficients	Estimate	S.E
AR1	−0.2006	0.1082
AR2	−0.2789	0.1168
AR3	−0.2194	0.1157
AR4	−0.4098	0.1197
AR5	−0.4332	0.1263

estimated model performed on the stationary series of total deceased COVID-19 cases dataset.

3.7 Classifiers

The Explorer module of WEKA was used to process the dataset. WEKA supports a variety of file formats, notably arff, CSV, JSON files, and so on. The dataset used is already in the CSV format so it can be simply imported by selecting the open file option and looking for the dataset's location. The Preprocess tab in the WEKA Explorer will show the dataset's current relation, attributes, instances, sum of weights, and visualisations once the dataset has been imported. Under the Classify tab, the user's options for machine learning algorithm classifiers will be shown when they click the choose button. There will be several folders with a list of the algorithms to be employed; Random Forest, is under the trees folder, SMOreg, Linear Regression, and Gaussian Processes are in the functions folder, and lBk is in the lazy folder. The test option needs to be filled out after choosing the appropriate algorithm; in this study, the researchers used 90/10 percentage split. The training set consists of times series data of total deceased cases in India from November 1, 2020, through January 6, 2021, while the testing set consists of the same time series data recorded from January 7 to January 13, 2021. The class attribute, which is immediately beneath the test choices section in this instance was "Total Deceased". Finally, the start button must be pressed to initiate the training process.

The results of the classifiers and ARIMA model were assessed based on their ability to predict the data present in the testing dataset. The experiment results of the same is illustrated in Table 4 and Fig. 8. The data in Table 4 and graphs in Fig. 8 give us an insight on the forecasting ability of the algorithms. To get a comprehensive insight on the predictive ability, we calculated the prediction error rate of the algorithms (see Table 5 and Fig. 9). From Tables 4 and 5, we surmise that ARIMA(5,2,0) model fits the actual values well and has the least overall prediction error rate ranging from 0.004 to 0.11%. We used two parameters, which included RMSE and MAE to measure their performance. Comparing the findings displayed in Table 6, we see that ARIMA(5,2,0) model performed the best as it has the lowest RMSE and MAE values. Thus, ARIMA(5,2,0) should be able to predict mortality in India in the coming weeks precisely.

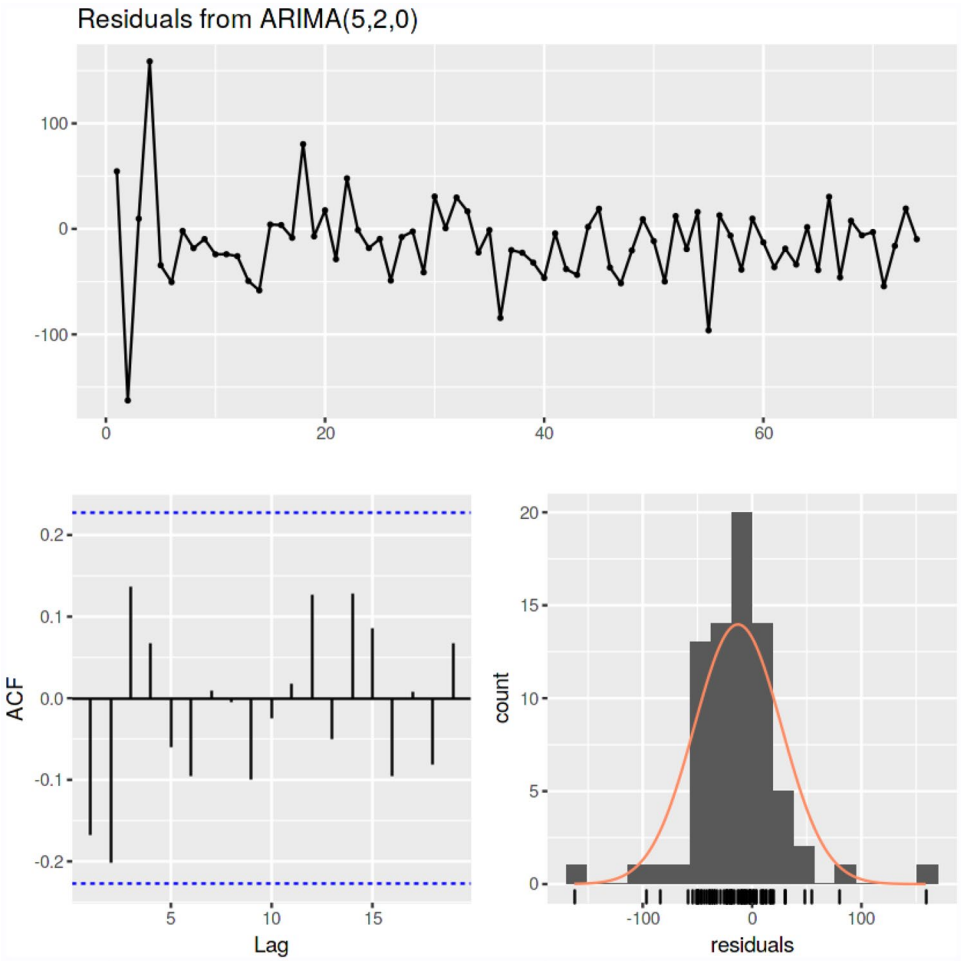

Fig. 7 Residual plots of ARIMA(5,2,0)

Table 4 Comparison of the predicted and actual values of ARIMA, SMOreg, Random Forest, lBk, Linear Regression and Gaussian Processes

Date	Actual	ARIMA	SMOreg	Random forest	lBk	Linear regression	Gaussian process
2021-01-07	150,015	150,007.0	149,944.5	150,034.7	149,898.5	149,848.6	151,853.1
2021-01-08	150,244	150,236.4	150,137.2	150,306.0	150,191.7	150,009.6	152,724.6
2021-01-09	150,457	150,450.9	150,326.7	150,559.1	150,360.2	150,194.8	153,619.5
2021-01-10	150,607	150,659.3	150,506.8	150,693.7	150,457.0	150,354.5	154,532.3
2021-01-11	150,773	150,887.9	150,660.6	150,795.0	150,690.0	150,495.2	155,463.3
2021-01-12	150,973	151,114.2	150,812.2	150,846.9	150,773.0	150,644.4	156,407.5
2021-01-13	151,174	151,341.4	150,969.5	150,971.3	150,973.0	150,771.8	157,385.2

4 Result and discussion

In the error evaluation, the findings were presented that the ARIMA(5,2,0) is better, where the RMSE 41.75084 and the MAE is 29.02792 as shown in Table 6. While, the RMSE values for Random Forest, lBk, SMOReg, Linear Regression, Gaussian Process are 43.547, 46.7127, 47.812, 52.8659, and 81.1246 respectively. The well-performed algorithm ARIMA(5,2,0) has been chosen to predict the rise in mortality due to COVID-19 in India. Figure 9 and Table 7 show the predicted total deceased cases from January 14, 2021 to February 12, 2021 (Fig. 10). The forecasted curve with 95% confidence level based on ARIMA(5,2,0) model (in blue) of fatality cases of COVID-19 for the next month and the present total deceased cases from November 11, 2020 to January 13, 2021 (in black) are shown in Fig. 11. Using the prediction of ARIMA model, the estimated daily average increase in number of total deceased cases is about 190 cases with a growth rate of 0.12–0.2

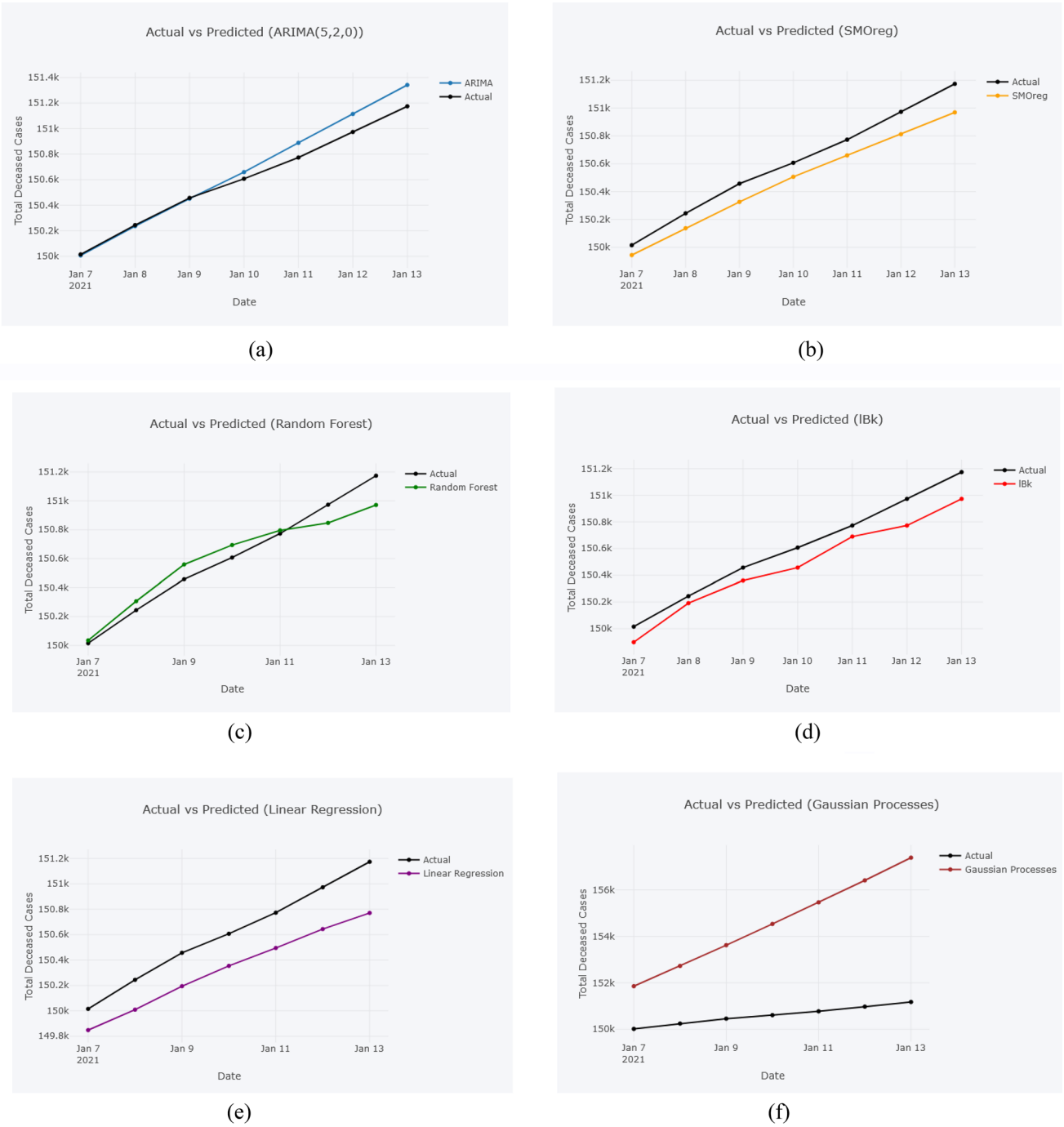

Fig. 8 Comparison of the predicted and actual values of **a** ARIMA, **b** SMOreg, **c** Random Forest, **d** lBk, **e** Linear Regression and **f** Gaussian Processes

every day. We can see clearly that the mortality rate of COVID-19 continues to increase slowly.

SARS-COV-2 is an infectious disease caused by a new coronavirus strain and has never been found in the human population before. This work emphasizes the potential of SMOreg, lBk, Random Forest, Linear Regression, Gaussian Processes and ARIMA methods to be used for forecasting total deceased cases of COVID-19. These methods are able to encapsulate time-variant properties and patterns of past data and forecast the future trend of COVID-19 time-series data. The estimated results revealed that the ARIMA model achieved higher accuracy compared to the other algorithms used for forecasting. Generally, there is no obvious answer that one of them is better than the other, but here the ARIMA

Table 5 Comparison of the prediction error rate of ARIMA, SMOreg, Random Forest, lBk, Linear Regression and Gaussian Processes

Date	ARIMA	SMOreg	Random forest	lBk	Linear regression	Gaussian process
2021-01-07	0.00532036	0.04699530	0.01312575	0.07765890	0.1109222	1.225277
2021-01-08	0.00503543	0.07108437	0.04123552	0.03484039	0.1560129	1.651048
2021-01-09	0.00408343	0.08660282	0.06784330	0.06436776	0.1742691	2.101929
2021-01-10	0.03469424	0.06653077	0.05757880	0.09959696	0.1676549	2.606320
2021-01-11	0.07623477	0.07454916	0.01459359	0.05504964	0.1842505	3.110835
2021-01-12	0.09349653	0.10650911	0.08355302	0.13247402	0.2176548	3.599650
2021-01-13	0.11071008	0.13527458	0.13410494	0.13295937	0.2660510	4.108643

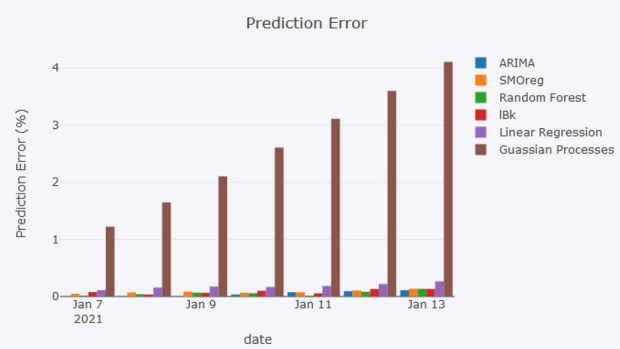

Fig. 9 Comparison of the prediction error rate of ARIMA, SMOreg, Random Forest, lBk, Linear Regression and Gaussian Processes

model outperforms the other forecasting approaches used for the COVID-19 prediction.

We reckon that statistical modeling assists in analyzing the outbreak and therefore we can foresee the course the pandemic is likely to take. Such a statistical model provides a broad view of the current pandemic and aids the authorities to work towards productive policies and decisions to limit impact of epidemic on society, medical care, and economic systems. In light of our outcomes, the predictors of the ARIMA model (with RMSE and MAE values of 41.75084 and 29.02792, respectively) performed better than other algorithms discussed above. In this way, the total number of deceased cases of COVID-19 in India was predicted using ARIMA(5,2,0) model. The findings uncover that mortality in India could rise from 151,174 to 157,179 in the next month with an average of 190 demise daily.

Developing a heuristic model to forecast COVID-19 deceased cases has its limitations starting from the dynamic nature of the disease and the accessibility of the data. The forcast model depends on the observation at a specific time, which is anticipated to be altered. Albeit the ARIMA model is normally a good solution to such a problem, there are few downsides to this methodology, it does not have innate updates, a new run has to be executed every time some new data is added. Also, the precision of the prediction depends on the amount of data.

5 Conclusion

In this work, we studied the pattern of the COVID-19 outbreak in India. Machine learning classification algorithms including SMOreg, lBk, Random Forest, Linear Regression and Gaussian Processes, along with ARIMA model have been applied to the real-time forecasts of the total deceased COVID-19 cases in India. We based our choice of algorithms used considering their extended capacity in capturing process nonlinearity and their versatility in modeling time-dependent data. Thirty days-ahead forecasts are provided based on historical data of 74 days since November 1, 2020 for India. RMSE, MAE and MAPE parameters were used to verify each model.

We discovered that the best predictor model for foreseeing the mortality in India is ARIMA (5,2,0). This model gave us the option to gauge the propagation rate of deceased cases for the next month. Accurately predicting the number of deceased cases provides information to governments and decision-makers about the expected situation and the necessary enforcement measures. Also, forecasting information can help encourage the public to contemplate the imposed measures to minimize the spread of this virus. The latest

Table 6 Performance evaluation of ARIMA, SMOreg, Random Forest, lBk, Linear Regression and Gaussian Processes based on RMSE and MAE values

Parameters	ARIMA (5,2,0)	SMOreg	Random forest	lBk	Gaussian processes	Linear regression
RMSE	41.75084	47.812	43.547	46.7127	81.1246	52.8659
MAE	29.02792	40.0454	38.6862	42.1429	58.4571	39.4917

Table 7 Forecast of total deceased COVID-19 cases in India for the next 30 days with 95% confidence level (C.L)

Date	Forecast	Low 95% C.L	High 95% C.L	Date	Forecast	Low 95% C.L	High 95% C.L
2021-01-14	151,395	151,308.6	151,480.6	2021-01-29	154,389	152,832.3	155,946.6
2021-01-15	151,624	151,447.1	151,801.2	2021-01-30	154,595	152,909.2	156,282.6
2021-01-16	151,825	151,556.1	152,094.8	2021-01-31	154,797	152,975.7	156,620.0
2021-01-17	152,010	151,648.9	152,372.0	2021-02-01	154,994	153,033.5	156,954.4
2021-01-18	152,196	151,753.3	152,639.1	2021-02-02	155,187	153,088.5	157,287.2
2021-01-19	152,380	151,870.7	152,890.0	2021-02-03	155,383	153,144.9	157,621.2
2021-01-20	152,576	151,995.8	153,155.9	2021-02-04	155,581	153,202.9	157,959.9
2021-01-21	152,788	152,126.3	153,450.4	2021-02-05	155,783	153,261.0	158,306.1
2021-01-22	153,001	152,243.9	153,758.8	2021-02-06	155,987	153,315.8	158,658.8
2021--01-23	153,207	152,340.1	154,074.6	2021-02-07	156,189	153,364.1	159,014.6
2021-01-24	153,406	152,422.5	154,391.3	2021-02-08	156,388	153,406.1	159,371.2
2021-01-25	153,597	152,498.4	154,697.0	2021-02-09	156,585	153,443.8	159,727.5
2021-01-26	153,786	152,575.0	154,997.1	2021-02-10	156,782	153,479.6	160,084.3
2021-01-27	153,981	152,659.1	155,303.7	2021-02-11	156,979	153,515.3	160,444.0
2021-01-28	154,183	152,747.4	155,619.9	2021-02-12	157,179	153,551.0	160,808.3

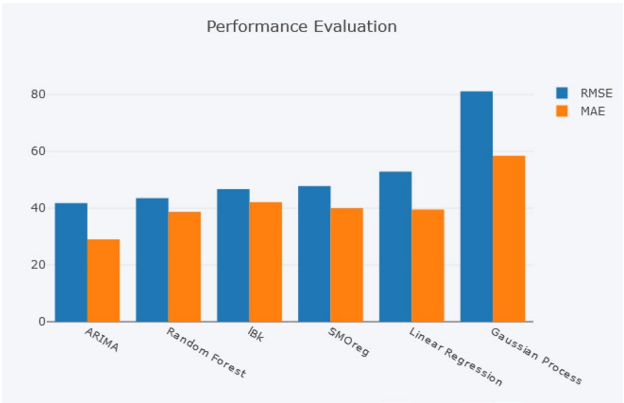

Fig. 10 Performance evaluation of ARIMA, SMOreg, Random Forest, lBk, Linear Regression and Gaussian Processes based on RMSE and MAE values

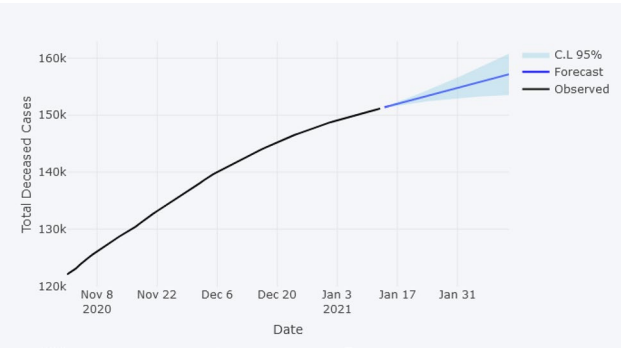

Fig. 11 Forecast of total deceased COVID-19 cases in India for the next 30 days with 95% confidence level (C.L)

trend shows a drop in COVID-19 cases but the pandemic is far from over. It simply means that the first phase of formidable challenges is over, and then the next phase begins. Vaccination will be a major addition to the tool kit we have. However, we don't know how long the vaccine protection will last. Solid requirement of social distancing measures, ceaseless monitoring and re-vaccination will still be needed to limit the spread of COVID-19.

Other models, such the SIR model or its variations, could be used in future research. These models could be compared in terms of precision and error magnitude. As in the case of the study's limitation, which cannot be addressed in the series of models described in this research, other aspects like meteorological and socio-economic effects must be considered. It could be interesting to observe how the epidemic evolves over the course of the summer and winter. The SARIMA model could be used to investigate the data's seasonality. The study's scope can be expanded by adding new data points and expanding it to other states and nations.

Author contribution VV: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Supervision, Writing—review & editing. AK: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing—original draft.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declarations

Conflict of 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.

References

Abolmaali S, Shirzaei S (2021) A comparative study of SIR model, linear regression, logistic function and ARIMA Model for forecasting COVID-19 cases. AIMS Public Health 8(4):598

Alabdulrazzaq H, Alenezi MN, Rawajfih Y, Alghannam BA, Al-Hassan AA, Al-Anzi FS (2021) On the accuracy of ARIMA based prediction of COVID-19 spread. Results Phys 27:104509

Alzahrani SI, Aljamaan IA, Al-Fakih EA (2020) Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions. J Infect Public Health 13(7):914–919

ArunKumar KE, Kalaga DV, Kumar CMS, Chilkoor G, Kawaji M, Brenza TM (2021) 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 103:107161

Barman A (2020) Time series analysis and forecasting of covid-19 cases using LSTM and ARIMA models. Preprint arXiv:2006.13852.

Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M (2020) Application of the ARIMA model on the COVID-2019 epidemic dataset. Data Brief 29:105340

Box GE, Jenkins GM, Reinsel GC, Ljung GM (2015) Time series analysis: forecasting and control. John Wiley & Sons, Hoboken

Ceylan Z (2020) Estimation of COVID-19 prevalence in Italy, Spain, and France. Sci Total Environ 729:138817

Coronaviruses NI (2020) National Institute of Allergy and Infectious Diseases. NIH Natl. Institue Allergy Infect. Dis. NIAID https://www.niaid.nih.gov/diseases-conditions/coronaviruses. Accessed 23 Aug 2021

De Gooijer JG, Hyndman RJ (2006) 25 years of time series forecasting. Int J Forecast 22(3):443–473

Dickey DA, Fuller WA (1981) Likelihood ratio statistics for autoregressive time series with a unit root. Econom J Econom Soc 49:1057–1072

Dimri T, Ahmad S, Sharif M (2020) Time series analysis of climate variables using seasonal ARIMA approach. J Earth Syst Sci 129(1):1–16

Faruk DÖ (2010) A hybrid neural network and ARIMA model for water quality time series prediction. Eng Appl Artif Intell 23(4):586–594

Fattah J, Ezzine L, Aman Z, El Moussami H, Lachhab A (2018) Forecasting of demand using ARIMA model. Int J Eng Bus Manag 10:1847979018808673

Ghosh S, Ghosh S (2020) Air quality during COVID-19 lockdown: blessing in disguise. Indian J Biochem Biophys 57:420–430

Gupta N, Tomar A, Kumar V (2020) The effect of COVID-19 lockdown on the air environment in India. Glob J Environ Sci Manag 6:31–40

Gupta R, Pal SK (2020) Trend analysis and forecasting of COVID-19 outbreak in India. medRxiv preprint https://doi.org/10.1101/2020.03.26.20044511

Hyndman RJ, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Wang E (2020) Package 'forecast' [Online]. https://cran.r-project.org/web/packages/forecast/forecast.pdf

Istaiteh O, Owais T, Al-Madi N, Abu-Soud S (2020) Machine learning approaches for COVID-19 forecasting. In: 2020 international conference on intelligent data science technologies and applications (IDSTA), pp. 50–57. IEEE

Jalloh MF, Li W, Bunnell RE, Ethier KA, O'Leary A, Hageman KM, Redd JT (2018) Impact of Ebola experiences and risk perceptions on mental health in Sierra Leone, July 2015. BMJ Glob Health 3(2):e000471

Konarasinghe KMUB (2021) SCM and SARIMA on forecasting COVID-19 outbreak in Italy. J New Front Healthc Biol Sci 2(1):20–38

Kumar A, Nayar KR (2021) COVID 19 and its mental health consequences. J Ment Health 30:1–2

Ljung GM, Box GE (1978) On a measure of lack of fit in time series models. Biometrika 65(2):297–303

Mahalakshmi G, Sridevi S, Rajaram S (2016) A survey on forecasting of time series data. In: 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), pp. 1–8. IEEE

Mahato S, Pal S, Ghosh KG (2020) Effect of lockdown amid COVID-19 pandemic on air quality of the megacity Delhi, India. Sci Total Environ 730:139086

Pandey G, Chaudhary P, Gupta R, Pal S (2020) SEIR and regression model based COVID-19 outbreak predictions in India. Preprint arXiv:2004.00958

Roy A, Singh AK, Mishra S, Chinnadurai A, Mitra A, Bakshi O (2021) Mental health implications of COVID-19 pandemic and its response in India. Int J Soc Psychiatry 67:587–600

Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52(3/4):591–611

Srivastava A (2021) COVID-19 and air pollution and meteorology-an intricate relationship: a review. Chemosphere 263:128297

Team RC (2013) A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online: www.r-project.org. Accessed 14 Feb 2019

Trapletti A, Hornik K, LeBaron B, Hornik MK (2020) Package 'tseries'. Available online: https://cran.r-project.org/web/packages/tseries/tseries.pdf

Vijayarani S, Suganya E, Navya C (2021) Crime analysis and prediction using enhanced Arima model. Journal homepage: www.ijrpr.com ISSN 2582, 7421

Villavicencio CN, Macrohon JJE, Inbaraj XA, Jeng JH, Hsieh JG (2021) COVID-19 Prediction applying supervised machine learning algorithms with comparative analysis using WEKA. Algorithms 14(7):201

Wang WC, Chau KW, Xu DM, Chen XY (2015) Improving forecasting accuracy of annual runoff time series using ARIMA based on EEMD decomposition. Water Resour Manag 29(8):2655–2675

Wei WWS (2013) Time series analysis. In: The Oxford handbook of quantitative methods in psychology: Vol. 2: statistical analysis, vol 2, pp. 458–485. Oxford University Press

Weka WEKA (2011) 3: data mining software in Java. University of Waikato, Hamilton, New Zealand (www.cs.waikato.ac.nz/ml/weka)

Wickham H, Averick M, Bryan J, Chang W, McGowan LDA, François R, Yutani H (2019) Welcome to the Tidyverse. J Open Source Softw 4(43):1686

World Health Organization (2008) HIV/AIDS and mental health report by the Secretariat. https://apps.who.int/gb/archive/pdf_files/EB124/B124_6-en.pdf

World Health Organization (2020) Coronavirus disease 2019 (COVID-19): situation report, 82. World Health Organization. https://apps.who.int/iris/handle/10665/331780

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.