IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 24, 2021, accepted June 30, 2021, date of publication July 5, 2021, date of current version July 15, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3094658

BIFM: Big-Data Driven Intelligent Forecasting Model for COVID-19

SUJATA DASH[1], (Member, IEEE), CHINMAY CHAKRABORTY[2], SOURAV KUMAR GIRI[1], SUBHENDU KUMAR PANI[3], AND JAROSLAV FRNDA[4], (Senior Member, IEEE)

[1]Department of Computer Application, Maharaja Sriram Chandra Bhanja Deo University, Baripada, Odisha 757003, India
[2]Department of Electronics and Communication Engineering, Birla Institute of Technology, Mesra, Ranchi, Jharkhand 835215, India
[3]Krupajal Computer Academy, Bhubaneswar, Odisha 751002, India
[4]Department of Quantitative Methods and Economic Informatics, Faculty of Operation and Economics of Transport and Communications, University of Žilina, 01026 Žilina, Slovakia

Corresponding author: Chinmay Chakraborty (cchakrabarty@bitmesra.ac.in)

This work was supported in part by the Operational Program Integrated Infrastructure for the Project (Identification and possibilities of implementation of new technological measures in transport to achieve safe mobility during a pandemic caused by COVID-19) through the ITMS Code under Grant 313011AUX5, and in part by the European Regional Development Fund.

ABSTRACT Ever since the pandemic of Coronavirus disease (COVID-19) emerged in Wuhan, China, it has been recognized as a global threat and several studies have been carried out nationally and globally to predict the outbreak with varying levels of dependability and accuracy. Also, the mobility restrictions have had a widespread impact on people's behavior such as fear of using public transportation (traveling with unknown passengers in the closed area). Securing an appropriate level of safety during the pandemic situation is a highly problematic issue that resulted from the transportation sector which has been hit hard by COVID-19. This paper focuses on developing an intelligent computing model for forecasting the outbreak of COVID-19. The autoregressive integrated moving average (ARIMA) machine learning model is used to develop the best model for twenty-one worst-affected states of India and six worst-hit countries of the world including India. The best ARIMA models are used for predicting the daily-confirmed cases for 90 days future values of six worst-hit countries of the world and six high incidence states of India. The goodness-of-fit measures for the model achieved 85% MAPE for all the countries and all states of India. The above computational analysis will be able to throw some light on the planning and management of healthcare systems and infrastructure.

INDEX TERMS ARIMA models, autocorrelation function, infectious disease, Ljung-box test, partial autocorrelation function, pandemics, time series models, white residual, transportation.

I. INTRODUCTION

In the end, The COVID-19 (SARS-CoV-2) pandemic poses an unprecedented threat to global public health. It has been reported as the most harmful contagious disease since the 1918 H1N1 influenza pandemic. According to the World Health Organization (WHO), COVID-19 situation report [1] as on March 31st, 2021, the pandemic has infected more than 128 million people worldwide with the USA reporting the highest number of cases (30,462,210), followed by Brazil (12,748,747), India (12,221,665), France (4,705,068), Russia (4,494,234) and United Kingdom(4,359,982) on the sixth position. The worldwide death toll stands at 2,815,939 with the highest number of deaths reported from the USA (552,352) followed by Brazil, Mexico, India, the UK, Italy while the number of recovered cases is 73,111,302. The disease has been spreading very aggressively, affecting most of the countries or territories globally. The SARS-CoV-2 virus belongs to the β-coronavirus family which is prevalent and has many possible natural hosts. This characteristic of the virus creates major hindrances for the prevention and cure of the infection. SARS-CoV-2 is highly infectious but low mortality rate [2] comparing with severe acute respiratory syndrome and Middle East respiratory syndrome coronaviruses (SARS-CoV and MERS-CoV) respectively. Another study from Peking University suggests that SARS-CoV-2 infection is in all likelihood caused by snakes [3], but it is later refuted by another study [4]. However, using gene-sequencing technology [5] a finding of the research from Wuhan Institute of Virology established a similarity of 96.2% between SARS-CoV-2 and bat coronavirus. However, the identification of infected people is very much important for the reduction

The associate editor coordinating the review of this manuscript and approving it for publication was Yuan Tian.

of virus spreading. People gathering at public transportation especially during rush hour (closed spaces with poor ventilation) can cause uncontrolled virus propagation. Due to this fact, an infected person should refrain from using public transportation.

Another study made by Xu *et al.* [6] using macro-genomic sequencing, molecular biological detection, and electron microscopic analysis achieved 99% similarity between SARS-CoV-2 strain taken from pangolins and the virus strains which is currently infecting humans.

COVID-19 is a highly infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and transmitted through respiratory droplets or with infected droplets. The coronavirus infections are infectious during the incubation period that may take 2 to 14 days to appear [7]; moreover, there is no approved vaccine or specialized medication available. The average number of people being contaminated by a single patient [8] varies from 1.5 – 3.5.

The novel coronavirus 2019 disease first emerged from the seafood markets of Wuhan, China on 31st December 2019. As of May 22, 2020, the number of infections in China stood at 82,971 with a death toll of over 4634. However, the overall case fatality rate in China diagnosed as of February 11, 2020 [9] is estimated to be 4.5% but in the age group of 70-80, it goes up to 8.0% while for above 80 it increases to14.8%. In addition, persons above the age of 50 with the ailments like diabetes, cardiovascular disease, and respiratory-related disease are more prone to become a victim of this disease.

The paper is organized as follows: Section I introduces the research work, Section II outlines the objectives and contributions of the study. Section III presents reviews of the literature, and Section IV presents the proposed models and methodologies to be used in this research. Section V explains the experimental techniques and analysis of the experimental findings. Section VI provides a discussion of the study followed by conclusion and limitations with future projection in Section VII and VIII respectively. Section IX represents an exhaustive bibliography.

II. OBJECTIVES & CONTRIBUTIONS OF THE STUDY

The objectives of the present study are as follows:

1) Developing the best ARIMA Models for twenty-one states of India and the top six countries of the world for evaluating the spread of the outbreak.
2) Studying the growth pattern and forecast of the outbreak for six most affected states of India and top six most affected countries including India employing ARIMA Model.
3) Studying the impact of lockdown on the incidence pattern of the disease only for *India*.

The contribution of the present study is outlined below which is obtained by achieving the above objectives. This paper focuses on an interesting aspect of the COVID-19 outbreak. The significant contributions are as follows:

1) The model provides an understanding of the number of people affected daily by this disease.
2) The forecasting of ARIMA models depicts 90 days future growth trend for confirmed cases for all six countries and six high incidence states of India.
3) The proposed model has achieved around 85% in terms of accuracy for all six countries and the six states of India.

This study will help to identify the factors which influence the growth pattern of incidence that may help to prioritize the challenges.

III. LITERATURE SURVEY

A student who returned from Wuhan, China on January 30, 2020, was identified as the first COVID patient in India and after that, no further cases were reported to increase the growth of the disease for a month [10] except three initial cases between January 30 and February 3, 2020. Since March 3, 2020, there has been a gradual increase in the number of infections until March 25, after that exponential rises of the cases are observed in different states of India. However, to contain and prevent the community transmission [11] of the disease, the Government of India enforced some early interventions like an international travel ban and a strict nationwide lockdown, limiting the movement of the 1.3 billion population of India. The phase 1 lockdown was imposed on 25 March for 21 days followed by phase 2 on 15 April, phase 3 on 4 May, and phase 4 on 18 May and phase 5 with a three-phase unlock plan that continued for the containment zone till 30th June 2020. The effect of lockdown is reflected in the spread of infections [11] across all the states of India which appears to be heterogeneous. Although the public health infrastructure of the country is inadequate to counter the enormity of the pandemic, the improvisations of diagnostic infrastructure have enabled the Indian Council of Medical Research (ICMR) [11] to increase its testing capacity to 3 lakh samples per day and further planning to scale up testing facility to tackle the migration of workers in states like Uttar Pradesh, Bihar, West Bengal, and Odisha. On the contrary, the mortality rate (14.27) which is still low in the world could be attributed greatly to the factors like hot and humid climate [12], high proportion of young people and BCG vaccinations [13] but these studies are in preliminary stages require more research to establish.

The forecasting of time series data predicts the future values employing mathematical and statistical techniques based on some specific assumptions considered for the underlying system [14], [15]. There are several predicting models available, each of them depends on a certain methodology and behave differently under different assumptions over the temporal evolution of the system [16]. Problems from various fields of science viz, information systems [17], electrical engineering [18], medical diagnosis applications [19]–[22], power management [23] and stock market [24] are addressed successfully by time series forecasting models. Again, this time series forecasting method can be split into two categories: short-term and long-term forecasting while short-term forecasting generates a robust prediction of even a few

hours ahead future values [25] by performing exhaustive analysis and computation of the underlying assumptions. On the other hand, long-term forecasts predict very long future values by analyzing the trend of the series and the parameters involved [26]. Hence, short-term predictions can be employed for clinical situations and long-term predictions assess patient's conditions even after many years.

In the past, the world has experienced several pandemics such as influenza, SARS, HIV/AIDS, and Ebola which originated in animals, induced by viruses [37], and assumed to evolve by socioeconomic fluctuations, ecological or environmental conditions. The last pandemic influenza 1 (H1N1) started in 2009 and escalated globally within a very short period. The resurgence of the disease [38], [60] was caused due to the antigenic drift and shifted to different parts of the world from time to time. Although, many of the scientific findings claim that COVID-19 is similar to a coronavirus family that originated from bats and the intervening host may be pangolin but it is still unresolved.

Several extensive works for prediction of the escalation of COVID-19 has been carried out using machine learning algorithms [28] such as neural networks for deep learning, polynomial fitting and exponential smoothing. Artificial intelligence (AI) have already proved their efficiency in predicting complex healthcare problems [29]–[31] like cancer, neurodegenerative disorders etc., with high precision. AI-driven methods [29] can be useful to predict the severity of the outbreak and can control the transmission of the disease. However, the neural network has encountered over-fitting problem due to small volume of data and polynomial fitting also faced the same problem with high bias [34]. The reason is because the trend of the spreading of the disease changes in different phases of the lockdown over time, these model shows variations in their pattern.

The study proposes to use ELM [54] with a fully connected layer to provide a real-time training phase. Besides, ELM's stochastic nature brings about an extra-uncertainty problem, particularly for high-dimensional image processing systems. Due to the stochastic choice of the input weights and biases in ELM, it leads to ill-conditioned matrices to system producing non-optimal solutions. To alleviate this issue, a novel meta-heuristic algorithm called Chimp Optimization Algorithm (ChOA) [54] to improve ELM conditioning and ensure optimal solutions is employed. Although different types of ELM [55]–[57] are now accessible for detecting image and classifying problems. Therefore, many researchers have worked extensively applying mathematical and statistical models [32] to understand the Spatio-temporal dynamics of COVID-19 outbreaks. These models gave a new impetus to understand the public policy for proper selection and allocation of resources and public health interventions [33] during the pandemics. Timely forecasts of the measures namely, peak time, peak height, and enormity during the pandemic would be beneficial for making reliable predictions for healthcare resources and manpower. When the capacity to develop, evaluate manufacture, distribute and administer effective medical countermeasures such as vaccines, diagnostics, therapeutics are inadequate to meet the burden of emerging outbreaks of infectious diseases, public health measures and supportive clinical care remain the only feasible tools to slow down the emerging outbreak. Under such circumstances, decision-making can be an alternative by the use of appropriate data and advanced analytics such as infectious disease modelling. Further new applications of data science and statistical analysis to disease outbreaks could provide support to decision-makers during a public health crisis. The models used to forecasts the trends of the outbreak and on which epidemiological stage the country is going through are based on regression models, Facebook Prophet, SEIR model, ARIMA model, prediction rules [34]–[37], etc.

The ARIMA model [41], [27] is a time series model, designed basically for economic applications. However, for the past few decades, it has been widely used by healthcare researchers to predict different aspects of infectious disease. Generally, the model removes the high-frequency noise present in the data, identifies the local trends based on linear dependence, and then forecasts the future trends [38]. Usually, time series models demonstrate high ability of prediction and extensive applicability than non-temporal methods [42]. So far as the use of the model is concerned, other than predicting the severity of the disease, it has been used to predict 3 days the number of hospital beds occupied and planning of other critical resources during the pandemic of severe acute respiratory syndrome (SARS) [43], [58], [59]. In addition, this analytical tool helps healthcare managers and researchers to measure the healthcare interventions in a specific population. Gupta and Pal [44] have used the ARIMA model to predict the number of infected cases in India for the best case, worst case, and average-case scenarios.

Even though the model displays high performance, still it has some constraints which curtail its range of applications such as:

1) Maintains a linear relationship between the dependent and predicting variables instead of reflecting the actual non-linear relationship exist in the dataset. 2) Assumes the mean and variance of the series is not time-dependent i.e., stationary [39]. 3) Assumes the residual time series follows Gaussian distribution. 4) Models are non-static and cannot be used for reconstructing the missing data.

ARIMA is a parametric method and it forecasts better for relatively short series when the number of observations are not adequate for applying advanced machine learning methods. In [50], six different statistical and machine learning-inspired time series models were developed for estimating the percentage of active cases for seven days ahead concerning the total population for the ten countries with the highest number of confirmed cases as of 4 May 2020. The comparison of the results of different approaches indicates that the traditional statistical methods namely, ARIMA and TBAT prevail over deep learning counterparts such as DeepAR and N-BEATS—an outcome which, due to the lack

of large amounts of data. As compared to other econometric models, ARIMA models have been used with success in the prediction of several diseases [51]–[53].

Hence, in this study, ARIMA model is adopted to understand the incidence and pattern of the trend of the pandemic coronavirus in the worst hit states of India and top six severely affected countries of the world. The countries and states of India have been selected based solely on the severity and high density of the infection in those locations as compared to others.

IV. PROPOSED METHOD DESCRIPTION & METHODOLOGY

The three main variables of this study are the number of confirmed cases, number of recoveries, and number of deaths. This study mainly focuses on developing the statistical regression models for forecasting the incidence, peak and trend of the outbreak.

A. PROPOSED AUTOREGRESSIVE INTEGRATED MOVING AVERAGE FORECASTING MODEL

ARIMA model can be applied to many real-time non-stationary time series problems like socio-economic, business and epidemiological studies [45] for prediction and analysis. In this paper, the non-seasonal ARIMA model is used to study the incidence of the COVID-19 pandemic in twenty-one severely affected states of India from 18th March 2020 to 31st March 2021 and worst hit top six countries including India for the period 30th January 2020 to 31st March 2021. This model consists of three parts: (i) an autoregressive part (AR), (ii) a contribution from a moving average (MA), and (iii) an integration part (I) and the model is denoted as ARIMA (p, d, q).

The autoregressive part AR (p) of the model is assumed to be a linear combination of the past p observations [38], and a random error with a constant term. The mathematical formulation is represented as:

$$\begin{aligned} y_t &= c + \sum_{i=1}^{p} \varphi_i y_{t-i} + \varepsilon_t \\ &= c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_p y_{t-p} + \varepsilon_t \end{aligned} \quad (1)$$

where y_t and ϵ_t are the target value and random error at period t, and $\varphi_i (i = 1, 2, \ldots, p)$ are the model parameters with a constant c. The integer constant p indicates the order of the model.

The moving average part MA (q) of the model uses past errors as the explanatory variables and mathematically represented as:

$$\begin{aligned} y_t &= \mu + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \\ &= \mu + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \end{aligned} \quad (2)$$

where μ is the mean of the series $\theta_j (j = 1, 2, \ldots, q)$ are the model parameters and q is the order of the model. The random error is a sequence of independent and identically distributed (i.i.d) random variables with zero mean and constant variance.

Integration of these two models develop ARIMA model and the mathematical formulation [43] is represented as follows:

$$y_t = \mu + \sum_{i=1}^{p} \varphi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \quad (3)$$

Here, the parameters are defined as:

- μ is the constant term, $p \geq 0$ is the order of AR model, AR(p) refers to the number of lags. $d \geq 0$ is the degree of differencing, I(d) refers to the integration parameter.
- $q \geq 0$ is the order of MA model, MA(q) refers to the number of lags.
- $\{\varphi_i\}, (i = 1, 2, \ldots, p)$ *are the parameters of AR(p) model*
- $\{\theta_j\}, (j = 1, 2, \ldots, q)$ *are the parameters of MA(q) model*
- ϵ_t *is the random error*

Differencing parameter is useful for increasing the stationarity of the series thereby reduces the mean to zero. Mathematically it can be represented as:

$$\Delta y_t = y_t - y_{t-1} \quad (4)$$

The parameter estimation, model building, and forecasting of the time series datasets consist of the following four phases of computation:

- ***Transformation phase:*** If the visualization of the time series datasets displays the characteristics of non-stationarity, then Kwiatkowski-Phillips-Schmidt-Shin (KPSS) a diagnostic tool [45] [61] can be applied for stationarity checking. Then the finite difference transformation technique defined in eq.(4) [38] makes necessary transformation in the series given in eq. (3) to produce the time series y_t confirming the characteristics of stationarity. After the transformation, the series is again tested with KPSS to check if the series is stationary around the mean or not.
- ***Model Identification stage:*** After obtaining the stationary series y_t from stage 1, the problem of best ARIMA(p, d, q) model identification at this stage determines the integer parameter p and q which governs the underlying process of y_t by examining the figures plotted by autocorrelation function (ACF) and partial autocorrelation function (PACF) [41] for the stationary series. This is the most crucial stage of the model building that involves fair amount of individual opinion regarding entertaining more than one structure of the model for further analysis. Thereby, this stage suggests further examination to narrow down the possible selection of best model from the candidate models of the series y_t.
- ***Best model estimation stage:*** This stage estimates all the candidate ARIMA models explored in the previous stage to obtain the best model of the series. For this

purpose, Box-Ljung test [46] can be applied for identifying the best-fitted model for the series while a conditional sum of square likelihood is used to estimate the model parameters. This test determines whether or not the errors are white noise or independent and identically distributed (i.i.d).

- ***Diagnostic checking stage:*** This diagnostic checking stage checks the adequacy of the model by examining the ACF and PACF of the residuals to ascertain that the autocorrelations of the error is very small and the model is a good fit of the series. Finally, the forecasting accuracy of the model is studied by calculating the root mean squared error (RMSE), mean absolute percentage error (MAPE), mean absolute error (MAE), and mean squared error (MSE).

The four stages explained above are depicted in the workflow diagram of the model shown in Figure 1.Then 80% of the diseased data is used to train the model while rest 20% data used for validating and predicting the future values of the pandemic after minimizing the bias and variance error. The model predicts 90 days' future values for the time-series datasets.

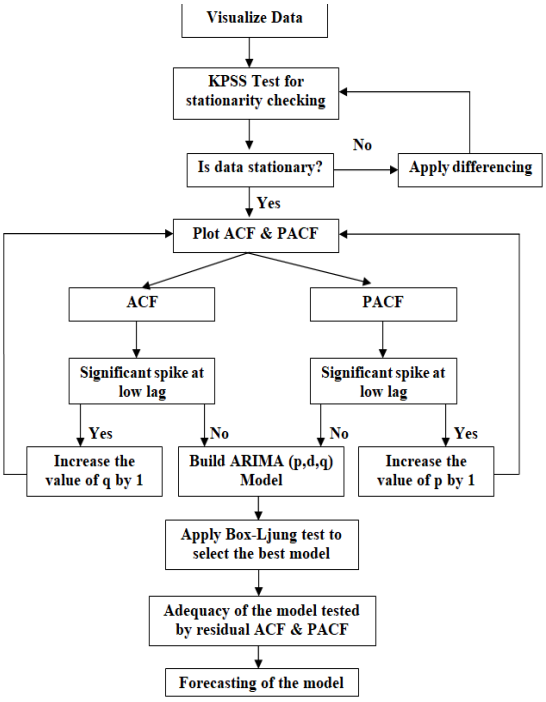

FIGURE 1. Proposed workflow diagram of ARIMA model.

V. EXPERIMENTAL SETUP AND ANALYSIS

All of the above methods are implemented using Python 3.8 version in Jupyter Notebook and executed in Windows 10, Intel(R), core-i7-7500U CPU @2.70GHz and 12.0 GB RAM. The packages which are used for prediction and visual representation of the findings are as follows: NumPy, pandas, SciPy, Matplotlib, ARIMA, sklearn, matplotlib, seaborn and statsmodels.

A. DATASET DESCRIPTION

All the datasets used in this analysis for predicting state wise disease status collected from the sources namely, https://www.covid19india.org/ [40] and for the six different countries collected from 'Johns Hopkins University Corona virus Data Stream that combines World Health Organization (WHO) and Centre for Disease Control and Prevention (CDC) case data'. The time-series datasets were imported using.CSV format.

B. EXPERIMENTAL RESULTS

The computation of the proposed method is carried out as follows:

- The computation pivots around the daily data of six different countries and twenty-one worst affected states of India to observe the changes in the trend of the incidence of the pandemic which is discussed in the following sub section. The four stages of the ARIMA model is applied to find the best-fitted ARIMA (p, d, q) model for the given countries and states.The computational procedure which is followed to build the best ARIMA (p, d, q) model is explicitly elaborated only for the time series data obtained for India in this section. For the remaining countries and states of India, only the best order model and the supporting Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) value, computed p-value from Box-Ljung test on residuals and predicted root-mean square error (RMSE) are given without computational details. Then the best models are used for forecasting the incidence of the disease for India and five countries of the world. Apart from this, the model of six most affected states of India is also used for forecasting the spread of the disease.

C. RESULTS OF BEST FITTED ARIMA(p,d,q) MODEL

Time series plots of confirmed, recovered and deceased data of India that display the observations on the y-axis against equally spaced time intervals on the x-axis used to evaluate the patterns and behaviors of the coronavirus disease over time is displayed in the Figure2.

1) ANALYSIS OF THE RESULT

The time series plots of the raw data collected for confirmed, recovered and death cases of India is displayed in Figure 2 which clearly shows a consistent increasing or decreasing patterns with outliers. This behavior reveals the inherent non-stationary characteristic of the data. The stationarity checking methods like unit root test determines in a formal way whether or not the series is stationary. A unit root test method namely, Kwiatkowski-Phillips-Schmidt-Shin (KPSS) [45] is employed on the series data whose result is presented in Table 1. The p-values of the KPSS test statistic is less than the critical value 0.05 which rejects the null hypothesis; 'H0: The series is stationary' at 5% level of significance. Hence, Figure 2 and Table 1 confirm that the time series data needs transformation to stationary or stabilize

TABLE 1. Stationarity testing of the time series data using KPSS.

Time Series Data	Test Statistics	Lag Order	p-value	Stationary (Y/N)
Confirmed	0.2494	16	0.01	No
Recovered	0.2046	16	0.01	No
Death	0.3082	16	0.01	No

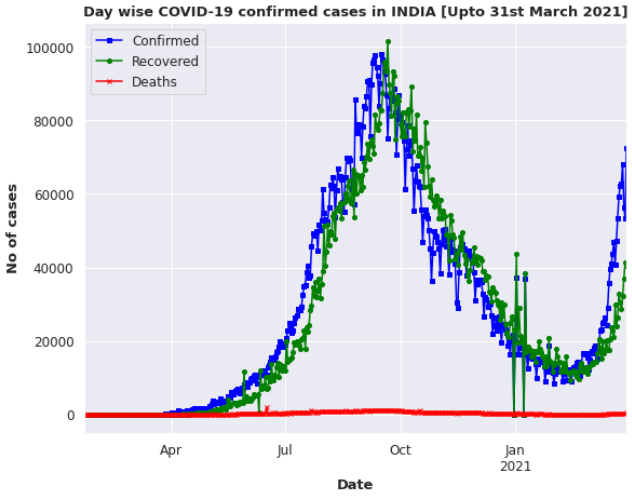

FIGURE 2. Time series graph of COVID-19 cases in India.

the series before being used for assessing the capability or initiating further improvements.

The time-series datasets with first-order differencing for recovered and deceased series and second-order differencing for confirmed series shown in Figure3 ascertain that all the three series achieve stationarity with zero mean and constant variance. The formal stationarity test method KPSS is applied to the first-order and second-order differenced series to obtain the statistical inference which is recorded in Table 2. The test statistic results of KPSS show that the p-values are greater than the critical value 0.05that rejects the null hypothesis; 'H0: The series is not stationary' and confirms achieving stationarity by the series except the confirmed series. Hence, the results of the test statistic substantiate that all the series with first or second order differencing have attained the stationarity.

TABLE 2. KPSS test to check stationarity of differenced data.

Time Series Data	Difference Order	Test Statistics	Lag Order	p-value	Stationary (Y/N)
Confirmed	2	0.0381	16	0.1	Yes
Recovered	1	0.0287	16	0.1	Yes
Death	1	0.0456	16	0.1	Yes

The ACF and PACF plots displayed in Figure 4 shows the first-order differenced data for recovered and deceased cases and second-order differenced data for confirmed cases. The analysis of ACF and PACF is necessary for determining a fitted model for a given time series data and the computed

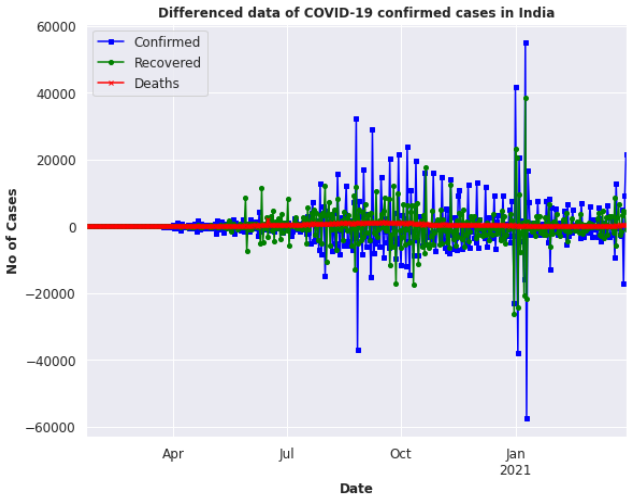

FIGURE 3. Time series graph of differenced COVID-19 data.

statistical measures shown in Table 2 illustrate the relationship between the observations in a time series data. More importantly, these plots determine the order of the AR and MA terms following some rules [47].

Applying the rules [47] on Figure 4 many feasible models can be computed for confirmed, recovered and deceased cases. The best model can be determined by finding the least p-value, RMSE, AIC and BIC. The estimates of the ARIMA(2,2,2) model is statistically significant as the p-value is less than 0.05 and the statistical AIC (8464.397), BIC (8488.821) and RMSE (5260.502) valuesare lower than the values of other models. Therefore, for confirmed time series dataset, ARIMA(2,2,2) is chosen as the best model for India.

For the recovered time series dataset ARIMA (1, 1, 1) is considered as the best model as it has obtained least value for statistical parameters viz, RMSE value (2169.498), AIC (2278.089) and BIC (2289.529). Applying the same rule ARIMA (2, 1, 0) model is statistically significant with parameters AIC (3244.521), BIC (3258.655) and RMSE (101.365), therefore, chosen as the best fitted modelfor deceased cases. All the estimated values of the parameters and coefficients of the three best-fitted models chosen for the confirmed, recovered and deceased time-series datasets of India are summarized in Table 3.

Figure 5(a, b, c) displays the line plot of residual errors of all three datasets, suggesting that the models have captured the trend information adequately. Again Figure 5(d, e, f) shows the density plot of the residual errors of confirmed, recovered and deceased models, suggesting the errors are Gaussian with little skewness. Then these fitted models are tested with Ljung-Box diagnostic tool where the p-value found to be very less than the usually chosen critical level of 0.05 that is shown in Table 4, in consequence the test is highly significant and therefore the null hypothesis is rejected, thus the residuals appear to be uncorrelated. This suggests that the residuals of the best fitted ARIMA(2,2,2), ARIMA(1,1,1) and ARIMA(2,1,0) models are white noise,

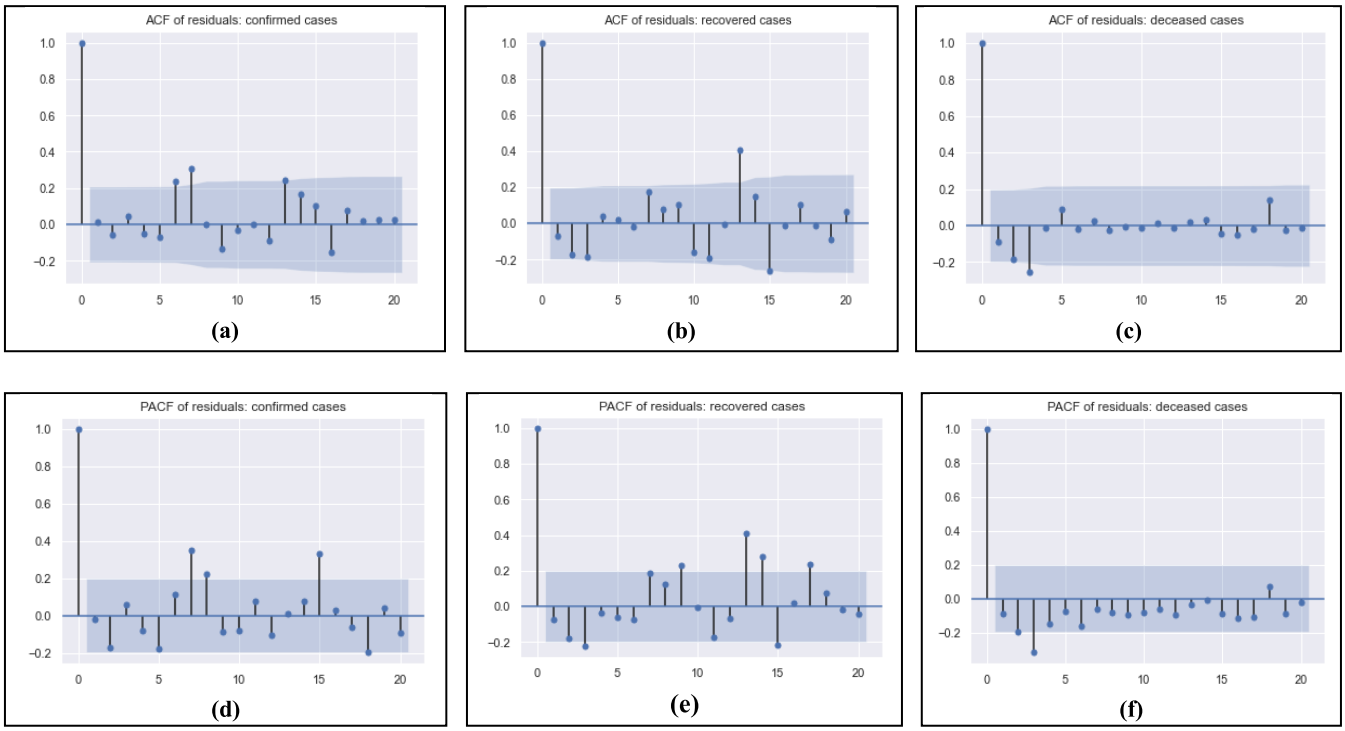

FIGURE 4. ACF and PACF of differenced data.

TABLE 3. Results of best ARIMA models.

Best Model	Parameter estimates	Coefficient	std error	p-value	z-value	AIC	BIC	RMSE
Confirmed Case ARIMA(2,2,2)	const	5.6579	10.216	0.580	0.554	8464.397	8488.821	5260.502
	AR1	0.2622	0.058	0.000	4.557			
	AR2	-0.2117	0.053	0.000	-4.008			
	MA1	-1.7349	0.037	0.000	-46.844			
	MA2	0.7828	0.037	0.000	20.957			
Recovered Case ARIMA(1,1,1)	Constant	46.492	0.000	0.000	4.502	2278.089	2289.529	2169.498
	AR1	0.104	0.004	0.014	-2.500			
	MA1	0.083	0.000	0.000	-7.322			
Deceased Case ARIMA(2,1,0)	Constant	1.8645	4.901	0.070	0.380	3244.521	3258.655	101.365
	AR1	-0.5800	0.060	0.000	-9.646			
	AR2	-0.2842	0.060	0.001	-4.738			

TABLE 4. Ljung-box test of best ARIMA models.

Type of Cases	Statistics	p-value
Confirmed	4880.38	< 0.05
Recovered	1006.4801	< 0.05
Deceased	200.3597	< 0.05

as a result, the models fit the series pretty well indicating that the parameters of the models are significant and the residuals are uncorrelated.

The plots in Fig. 7 comprising ACF and PACF plot of the residuals of the three models. The time plots of the residuals of the three best models clearly show that the residuals appear to be randomly scattered, no evidence of the correlation among the error terms. Therefore, the residual of errors are considered as an independent identically distributed (i.i.d) sequence with constant variance and zero mean. The ACF and PACF plots of the residuals of the three best models displayed in Fig. 7 shows some spikes at the lower lags of recovered model and one significant spike at 7^{th} lag of the confirmed model which can be ignored. The remaining spikes lies within the confidence boundary indicating that the residuals are most likely uncorrelated.

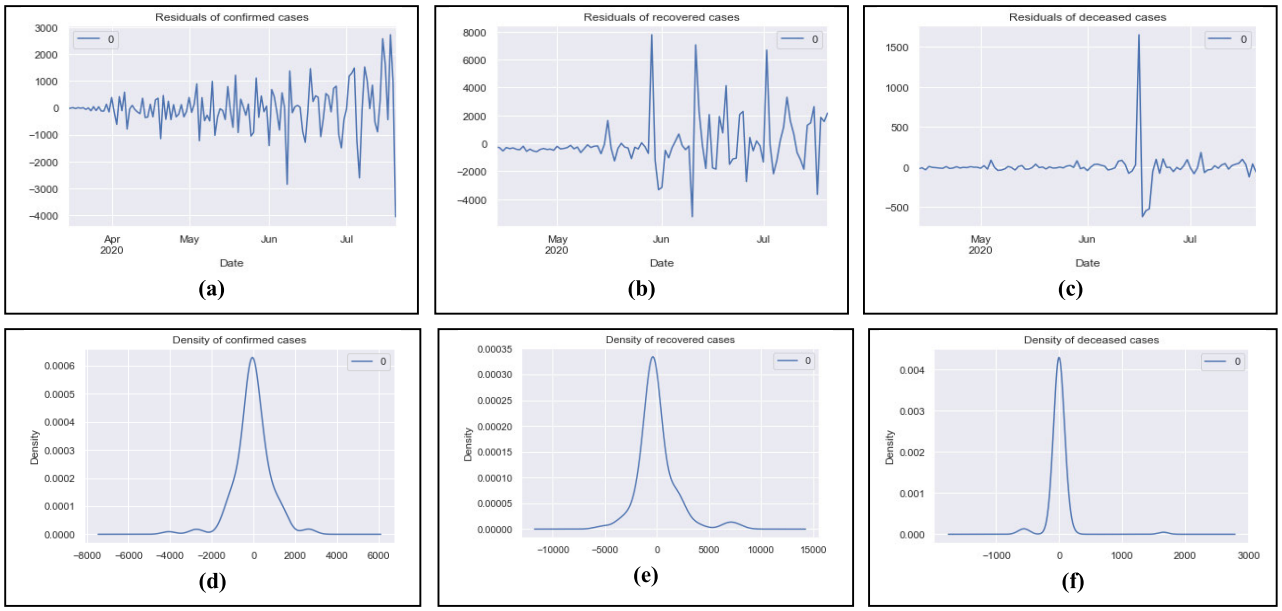

FIGURE 5. Residuals and density plots.

TABLE 5. Best fitted ARIMA model for top six worst affected countries (Confirmed cases).

Sl	Country Name	ARIMA order	p-value (Ljung-Box test)	AIC	BIC	RMSE Predicted
1	US	(0,1,0)	< 0.05	9751.353	9759.499	17491.17
2	Brazil	(2,1,2)	< 0.05	9403.873	9428.312	18813.737
3	India	(2,2,2)	< 0.05	8464.397	8488.821	5260.50
4	France	(2,1,2)	< 0.05	9156.750	9181.188	9870.182
5	Russia	(0,2,1)	< 0.05	6747.153	6759.366	565.891
6	United Kingdom	(2,1,2)	< 0.05	7978.710	8003.149	3603.14

For the remaining countries viz, USA, Brazil, France, Russia and UK only best ARIMA models for confirmed cases are given in Table 5 following the same experimental procedure as used for India. Table 5 records the best ARIMA model, computed p-value using Ljung Box test, AIC, BIC and predicted RMSE values for the above countries including India. The results established that the first order differencing is sufficient for the time-series data of USA, Brazil, Russia, France and UK to attain stationarity except India which needs second order differencing for attaining stationarity. Also the p-value is highly significant for all the models. As it is known that AIC and BIC are both penalized likelihood criteria and they try to balance good fit with parsimony. The lower the value of AIC and BIC the model is more closure to the true model. However, as the RMSE is scale dependent, one cannot assert a universal number as a good RMSE. As per our understanding, lower the values of RMSE better is the fitting and also it is a measure of how concentrated the data around the line of best fit. Therefore, based on the diagnostic metrics viz, computed p-value using Ljung Box test, AIC, BIC and predicted RMSE value, all the six models can be declared as the best fitted model.

Similarly, following the same computational procedure best ARIMA model for confirmed cases of 21 most affected states of India are designed. For all 21 states, computed p-value using Ljung-Box test, AIC, BIC and predicted RMSE values of the best ARIMA model are recorded in Table 6. All the models have realized least statistical Ljung-Box p-value, AIC, BIC and predicted RMSE to become the best fitted model. Only the confirmed series of Maharashtra and Gujarat have attained stationarity after second-order differencing. Only the RMSE value of Maharashtra (2226.635) is very high comparing with other states and the infection rate is also much higher. As a result, the higher AIC, BIC and predicted RMSE values of the state ensure over-fitting of the models. Also it is evident from the experiment that the residuals of the ARIMA model are not i.i.d. But for all other ARIMA models of the remaining twenty states the residuals of the models are white noise. Thus the residual plots corroborate the diagnostics that made by Ljung-Box test and the models fit well to the true model.

2) FORECASTING OF ARIMA MODELS

In the time series modeling, researchers expect to forecast the future values with minimum errors. In this section,

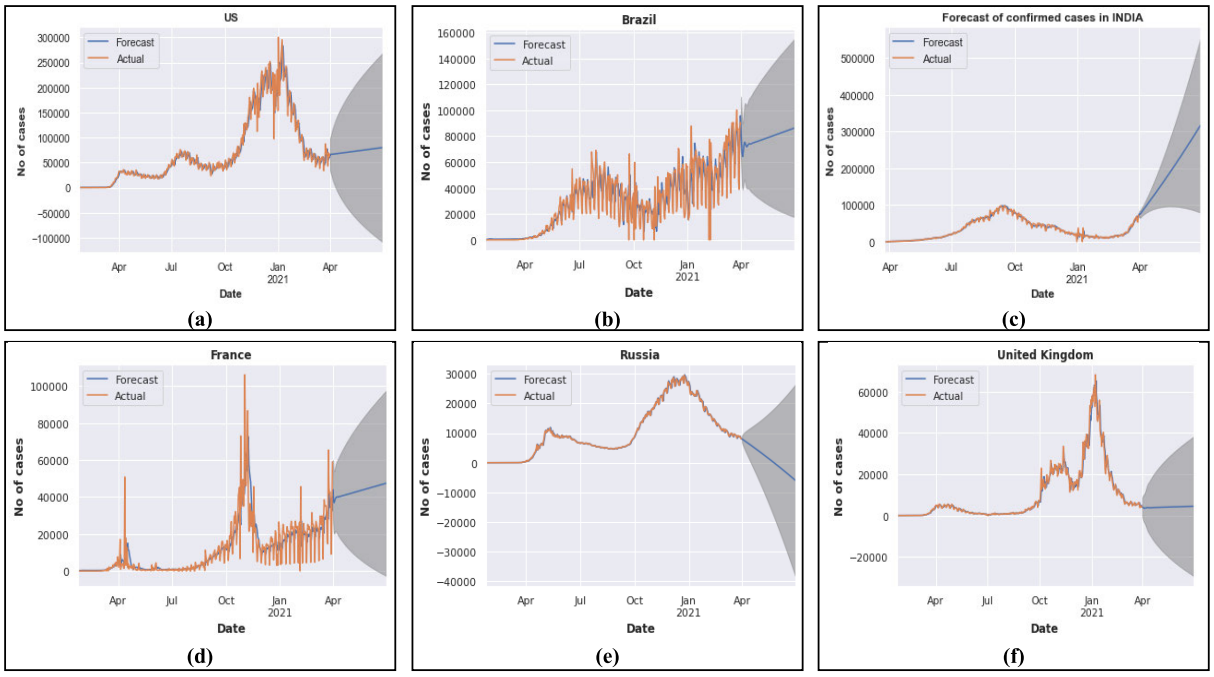

FIGURE 6. Forecasting trend of confirmed cases of six countries using ARIMA models.

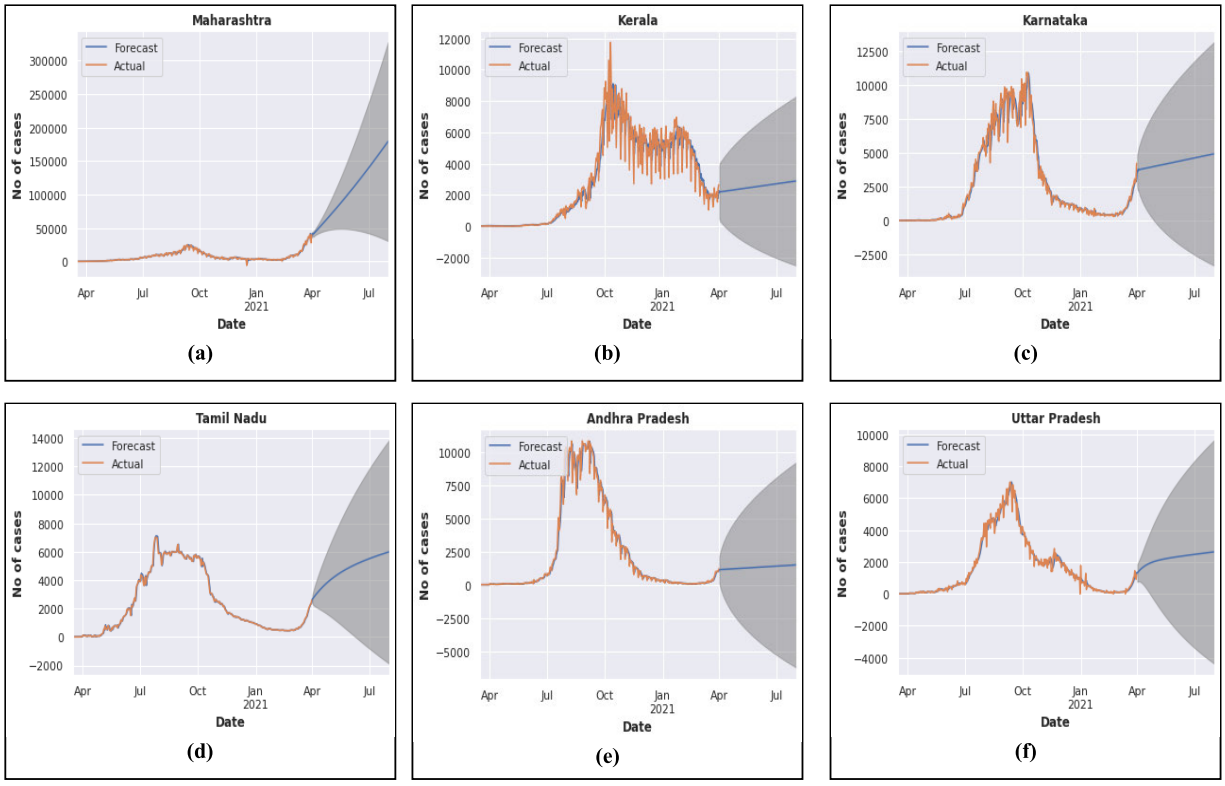

FIGURE 7. Forecasting trend of six most affected States of India using ARIMA models.

the forecasting performances of the ARIMA models of the confirmed cases of six most affected countries are discussed. The fitted ARIMA models forecast the outbreak for 90 days ahead with 95% confidence interval i.e. lower and upper confidence boundary for 90 days from 31st March 2021.

Figure 6 depicts the observed and predicted plots of confirmed coronavirus cases of six worst hit countries. The forecast of USA, Brazil, France, and UK show that the incidence of the disease will increase moderately for the next 90 days after 31st March 2021 whereas for Russia it declines steadily.

TABLE 6. Best fitted ARIMA model for most affected Twenty-One states of India (Confirmed cases).

Sl No	State Name	ARIMA	p-value	AIC	BIC	RMSE
1	Maharashtra (MH)	(2,2,2)	< 0.05	6730.263	6753.919	2226.635
2	Kerala (KL)	(1,1,2)	< 0.05	6256.955	6276.682	924.891
3	Karnataka (KA)	(2,1,0)	< 0.05	5989.567	6005.348	207.931
4	Tamil Nadu (TN)	(2,1,2)	< 0.05	4891.628	4915.301	40.593
5	Andhra Pradesh (AP)	(0,1,1)	< 0.05	5861.283	5873.119	81.645
6	Uttar Pradesh (UP)	(1,1,2)	< 0.05	5407.392	5427.119	205.129
7	Delhi (DL)	(2,1,0)	< 0.05	5872.965	5888.747	244.951
8	West Bengal (WB)	(0,1,1)	< 0.05	4789.561	4801.398	169.840
9	Chhattisgarh (CT)	(0,1,1)	< 0.05	5760.877	5772.713	738.182
10	Rajasthan (RJ)	(2,1,1)	< 0.05	4420.809	4440.536	89.425
11	Madhya Pradesh (MP)	(0,1,0)	< 0.05	5477.492	5485.383	70.852
12	Gujarat (GJ)	(1,2,0)	< 0.05	4457.400	4469.228	43.927
13	Odisha (OR)	(2,1,2)	< 0.05	4906.929	4930.602	30.735
14	Haryana (HR)	(0,1,0)	< 0.05	4791.296	4799.187	72.561
15	Telangana (TG)	(0,1,1)	< 0.05	5267.136	5278.972	77.665
16	Bihar (BR)	(2,1,1)	< 0.05	5195.732	5215.459	75.459
17	Punjab (PB)	(2,1,1)	< 0.05	4894.902	4914.629	151.514
18	Assam (AS)	(0,1,1)	< 0.05	5547.799	5559.635	15.796
19	Jharkhand (JH)	(0,1,0)	< 0.05	5139.512	5147.403	48.631
20	Jammu and Kashmir(JK)	(1,1,2)	< 0.05	4531.150	4550.877	41.172
21	Uttarakhand (UT)	(0,1,1)	< 0.05	4826.059	4837.895	78.530

The forecasting is depicted in Figure 6 (a-f). The predicted curve of Figure 6 (a, b, d, f) shows the per day infection cases of USA, Brazil, France, and UK which will reach around 75000, 85000, 50000 and 5000 respectively by the end of June 2021. Figure 6(e) shows, Russia will reach the ground by the end of June 2021. Figure 6(c) shows a sharp increase in the number of confirmed cases of India from April 2021 and the number of cases will be more than 3lakhs per day by the end of June 2021.

For analyzing the trend of the pandemic only the best models of six states of India are considered for further analysis and prediction. Therefore, out of the 21 developed ARIMA models, only first 6 models are considered here for forecasting purpose and the plots are shown in Figure 7(a-f). The states of Kerala, Karnataka, Tamil Nadu, Andhra Pradesh and Uttar Pradesh show a moderate growth in the trend of the disease Figure 7(b, c, d, e, f) whereas the forecast of Maharashtra Figure 7 (a) shows a steep rise in the growth curve. The predicted curve of Figure 7(a, b, c, d, e, f) shows the per day infection cases which will reach around 1.7lakhs, 3000, 5000, 6000, 1800 and 2500 for Maharashtra, Kerala, Karnataka, Tamil Nadu, Andhra Pradesh and Uttar Pradesh by the end of June 2021.

Finally, discussing about the estimated value of the performance metrics of ARIMA which is recorded in Table 7 establish that ARIMA model is robust in forecasting the future trend of confirmed cases of the pandemic. The RMSE and MAPE values of the ARIMA models are very significant for all the countries. However, the accuracy is more than 86% for USA, 72% for Brazil, 86% for India, 87% for France, 97% for Russia and 89% for UK. This confirms the efficiency and efficacy of ARIMA Model for being used as an epidemiological model to study the incidence of the disease.

VI. DISCUSSION

Several researchers have studied the incidence of COVID-19 and forecasted the spread of the disease for various countries and provinces. Suitable forecasting models capture the information from the time-series data thoroughly and provide better understanding of the spread of the disease across the population which helps to decide pertinent measures to control the transmission of the infection and increase the capacity of the healthcare system. At this moment, epidemiological solutions are highly essential rather than pharmaceutical solutions. However, it is crucial to assess the efficiency of the applied interventions for taking timely actions to alleviate the pandemic. These timely actions need precise information about the ongoing disease, accurate growth predictions, and reassessment of the implemented interventions. The present study endeavored to forecast the current scenario using regression models considering the data from 30th January 2020 to 31st March 2021 for five countries and 18th March 2020 to 31st March 2021 for India which projected the daily cases very close to the observed cases. ARIMA Model

TABLE 7. Performance metrics of ARIMA models.

Country	MSE	RMSE	MAE	MAPE %
US	305941199.44	17491.17	13086.25	13.81
Brazil	353956688.91	18813.73	14012.08	28.30
India	27672884.88	5260.50	3034.57	14.30
France	97420491.33	9870.18	7156.38	13.00
Russia	320233.13	565.89	430.78	3.04
UK	12982626.86	3603.14	2211.59	11.06

forecast 90 days ahead future values of the daily growth of the confirmed cases for top six most affected countries including India and six worst hit states of India. This model is based on the time-series data of the confirmed cases and forecasts future cases.A study by Hyndman and Khandakar [26] predicted using ARIMA model that the expected number of cases in the worst case may increase up to 700,000 and in an average case may increase up to 7000 but can be restricted to 1000 cases by 24th April if very strict measures are taken. However, the daily cases added on 24th April were 1408 and the total cases were 24,447 which contradicts the predictions. Another study by Khan and Gupta [27] using ARIMA model predicted 50 days daily cases which will go up to 1500 by the end of 20th March 2020 but the actual per day cases were only 55 and the total cases were 249 which indicates a gross underestimation. The ARIMA models developed in this paper predicted 90 days of future incidence from 31st March 2021 and the pattern of growth of the disease. The model shows India will enter to a second wave at the end of March with a much higher incidence rate. The prediction curve of Kerala and Andhra Pradesh show an overfitting problem thereby the forecasting trend may not reflect the future scenario. The remaining Indian states will enter a second wave most likely after March 2021. Lockdown was one of the major interventions imposed by the Government relatively bit early along with other public health precautions to alleviate the transmission of the pandemic. It raises an apparent question on the effectiveness of lockdown over the incidence cases. The effectiveness of the interventions [48], [49] is measured by many studies with varying levels of outcomes. The prediction models of France and Russia indicate a second wave with much severity. However, as discussed above necessitates the revision of the forecast model in a regular interval as and when the disease data gets available. Cross-country performance was hard to explain and interpret. However, important factors that should be noted include discrepancies among the different countries in terms of climatic and geographical characteristics; in terms of population-related characteristics such as density; in terms of COVID-19 measures and testing procedures; and terms of timing, duration, and severity of any social distancing measures if any that could be implemented will enhance the prediction in a more effective way.

VII. CONCLUSION

The best ARIMA models developed for six high incidence countries of the world and six severely affected states of India project 90 days ahead of future values to understand the spread of the pandemics. The model predicts very near to the exact values except for few exceptions. The ARIMA model selected for the UK is not the best fit. The predicted RMSE, AIC, and BIC values are reasonable for all six countries. Similarly, the predicted RMSE, AIC, and BIC values of ARIMA models designed for the states of India are quite significant. The statistical performance metrics prove the efficiency of the ARIMA Model. The findings of this model may be used for making plans for possible interventions to strengthen the healthcare system for better management of the infected people in India and other countries.

VIII. LIMITATIONS & FUTURE WORK

The main aim of this study is to forecast the trend pattern of the incidence of the disease. The limitations of the proposed models not only depend on the underlying assumptions but many factors viz, the density of population, healthcare system, interventions imposed by the administration, economic and socio-demographic situation. If the data on testing and screening strategies, policies adopted by different countries, information about the access of pre-exposure drug profile, and robustness of the healthcare system would be available for analyzing the existing information, one can incorporate these statistics to develop a robust predictive model. Also, the adoption of multivariate time series modeling that takes into account other factors that are either directly or indirectly related to the spread of the pandemic could predict more effectively. Another future ambition would be to use some form of transfer learning to bring learnings from one country to another.

REFERENCES

[1] *Coronavirus Disease 2019 (COVID-19): Situation Report, 123*, World Health Org., Geneva, Switzerland, 2020.
[2] L. Garg, E. Chukwu, N. Nasser, C. Chakraborty, and G. Garg, "Anonymity preserving IoT-based COVID-19 and other infectious disease contact tracing model," *IEEE Access*, vol. 8, pp. 159402–159414, 2020, doi: 10.1109/ACCESS.2020.3020513.
[3] W. Ji, W. Wang, X. Zhao, J. Zai, and X. Li, "Cross-species transmission of the newly identified coronavirus 2019-nCoV," *J. Med. Virol.*, vol. 92, no. 4, pp. 433–440, Apr. 2020, doi: 10.1002/jmv.25682.
[4] C. Zhang, W. Zheng, X. Huang, E. W. Bell, X. Zhou, and Y. Zhang, "Protein structure and sequence reanalysis of 2019-nCoV genome refutes snakes as its intermediate host and the unique similarity between its spike protein insertions and HIV-1," *J. Proteome Res.*, vol. 19, no. 4, pp. 1351–1360, Apr. 2020, doi: 10.1021/acs.jproteome.0c00129.
[5] P. Zhou *et al.*, "A pneumonia outbreak associated with a new coronavirus of probable bat origin," *Nature*, vol. 579, no. 7798, pp. 270–273, Mar. 2020, doi: 10.1038/s41586-020-2012-7.

[6] X. Xu, P. Chen, J. Wang, J. Feng, H. Zhou, X. Li, W. Zhong, and P. Hao, "Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission," *Sci. China Life Sci.*, vol. 63, no. 3, pp. 457–460, Mar. 2020, doi: 10.1007/s11427-020-1637-5.
[7] T. Singhal, "A review of coronavirus disease-2019 (COVID-19)," *Indian J. Pediatrics*, vol. 87, no. 4, pp. 1–6, 2020.
[8] Z. Wu and J. M. McGoogan, "Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: Summary of a report of 72 314 cases from the Chinese center for disease control and prevention," *JAMA*, vol. 323, no. 13, pp. 1239–1242, Apr. 2020, doi: 10.1001/jama.2020.2648.
[9] S. J. Kang and S. I. Jung, "Age-related morbidity and mortality among patients with COVID-19," *Infection Chemotherapy*, vol. 52, no. 2, pp. 154–164, 2020, doi: 10.3947/ic.2020.52.2.154.
[10] M. Rawat. India Today. (Mar. 12, 2020). *Coronavirus in India: Tracking Country's First 50 COVID-19 Cases; What Numbers Tell?*. [Online]. Available: https://www.indiatoday.in/india/story/coronavirus-in-india-tracking-country-s-first-50-covid-19-cases-what-numbers-tell-1654468-2020-03-12
[11] MOHFW. (Mar. 2020). *Enabling Delivery of Essential Health Services During the COVID 19 Outbreak: Guidance Note*. [Online]. Available: https://www.mohfw.gov.in/
[12] Y. Ma, Y. Zhao, J. Liu, X. He, B. Wang, S. Fu, J. Yan, J. Niu, J. Zhou, and B. Luo, "Effects of temperature variation and humidity on the death of COVID-19 in Wuhan," *China. Sci. Total Environ.*, vol. 724, Jul. 2020, Art. no. 138226, doi: 10.1016/j.scitotenv.2020.138226.
[13] S. Salman and M. Salem, "The mystery behind Childhood sparing by COVID-19," *Int. J. Cancer Biomed. Res.*, vol. 4, pp. 27–28, Jun. 2020, doi: 10.21608/JCBR.2020.26503.1022.
[14] A. A. Ariyo, A. O. Adewumi, and C. K. Ayo, "Stock price prediction using the ARIMA model," in *Proc. UKSim-AMSS 16th Int. Conf. Comput. Modeling Simul.*, Mar. 2014, pp. 26–28, doi: 10.1109/UKSim.2014.67.
[15] S. Aminikhanghahi and D. J. Cook, "A survey of methods for time series change point detection," *Knowl. Inf. Syst.*, vol. 51, no. 2, pp. 339–367, May 2017, doi: 10.1007/s10115-016-0987-z.
[16] M. P. Naeini, H. Taremian, and H. B. Hashemi, "Stock market value prediction using neural networks," in *Proc. Int. Conf. Comput. Inf. Syst. Ind. Manage. Appl. (CISIM)*, Oct. 2010, pp. 132–136.
[17] B. Spencer, O. Alfandi, and F. Al-Obeidat, "A refinement of lasso regression applied to temperature forecasting," *Procedia Comput. Sci.*, vol. 130, pp. 728–735, 2018, doi: 10.1016/j.procs.2018.04.127.
[18] M. Stanke and S. Waack, "Gene prediction with a hidden Markov model and a new intron submodel," *Bioinformatics*, vol. 19, no. 2, pp. ii215–ii225, Sep. 2003, doi: 10.1093/bioinformatics/btg1080.
[19] V. Ediger and S. Akar, "ARIMA forecasting of primary energy demand by fuel in turkey," *Energy Policy*, vol. 35, no. 3, pp. 1701–1708, Mar. 2007, doi: 10.1016/j.enpol.2006.05.009.
[20] New York, NY, USA. (Apr. 2019). *Yahoo Finance*. [Online]. Available: https://nance.yahoo.com/quote/NSRGY/history?p=NSRGY
[21] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 2nd ed. Melbourne, VIC, Australia: OTexts, 2018.
[22] E. Gijo, "Demand forecasting of tea by seasonal ARIMA model," *Int. J. Bus. Excellence*, vol. 4, no. 1, pp. 111–124, Jan. 2011, doi: 10.1504/IJBEX.2011.037252.
[23] K. B. Hemanta, C. Chinmay, K. P. Subhendu, and K. R. Vinayak, "Feature and sub-feature selection for classification using correlation coefficient and fuzzy model," *IEEE Trans. Eng. Manage.*, early access, Apr. 19, 2021, doi: 10.1109/TEM.2021.3065699.
[24] E. Guresen, G. Kayakutlu, and T. U. Daim, "Using artificial neural network models in stock market index prediction," *Expert Syst. Appl.*, vol. 38, no. 8, pp. 10389–10397, Aug. 2011, doi: 10.1016/j.eswa.2011.02.068.
[25] A. V. Metcalfe and P. S. P. Cowpertwait, *Introductory Time Series With R*. New York, NY, USA: Springer, Jun. 2009.
[26] R. J. Hyndman and Y. Khandakar, "Automatic time series forecasting: The forecast Package for R," *J. Stat. Softw.*, vol. 27, no. 3, pp. 1–22, 2008, doi: 10.18637/jss.v027.i03.
[27] F. M. Khan and R. Gupta, "ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in india," *J. Saf. Sci. Resilience*, vol. 1, no. 1, pp. 12–18, Sep. 2020.
[28] Q.-H. Ye, L.-X. Qin, M. Forgues, P. He, J. W. Kim, A. C. Peng, R. Simon, Y. Li, A. I. Robles, Y. Chen, Z.-C. Ma, Z.-Q. Wu, S.-L. Ye, Y.-K. Liu, Z.-Y. Tang, and X. W. Wang, "Predicting hepatitis b virus–positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning," *Nature Med.*, vol. 9, no. 4, pp. 416–423, Apr. 2003.
[29] M. V. Mai and M. Krauthammer, "Controlling testing volume for respiratory viruses using machine learning and text mining," in *Proc. AMIA Annu. Symp.*, 2016, pp. 1910–1919.
[30] G. Purcaro, C. A. Rees, W. F. Wieland-Alter, M. J. Schneider, X. Wang, P.-H. Stefanuto, P. F. Wright, R. I. Enelow, and J. E. Hill, "Volatile fingerprinting of human respiratory viruses from cell culture," *J. Breath Res.*, vol. 12, no. 2, Mar. 2018, Art. no. 026015, doi: 10.1088/1752-7163/aa9eef.
[31] E. O. Nsoesie, J. S. Brownstein, N. Ramakrishnan, and M. V. Marathe, "A systematic review of studies on forecasting the dynamics of influenza outbreaks," *Influenza Respiratory Viruses*, vol. 8, no. 3, pp. 309–316, 2014.
[32] S. Mandal, T. Bhatnagar, N. Arinaminpathy, A. Agarwal, A. Chowdhury, M. Murhekar, R. R. Gangakhedkar, and S. Sarkar, "Prudent public health intervention strategies to control the coronavirus disease 2019 transmission in India: A mathematical model-based approach," *Indian J. Med. Res.*, vol. 151, nos. 2–3, pp. 190–199, 2020.
[33] L. Jia, K. Li, Y. Jiang, X. Guo, and T. zhao, "Prediction and analysis of coronavirus disease 2019," 2020, *arXiv:2003.05447*. [Online]. Available: https://arxiv.org/abs/2003.05447
[34] R. Ranjan, "Predictions for COVID-19 outbreak in India using epidemiological models," *medRxiv*, doi: 10.1101/2020.04.02.20051466.
[35] G. Pandey, P. Chaudhary, R. Gupta, and S. Pal, "SEIR and regression model based COVID-19 outbreak predictions in india," 2020, *arXiv:2004.00958*. [Online]. Available: https://arxiv.org/abs/2004.00958
[36] S. S. Morse, J. A. Mazet, M. Woolhouse, C. R. Parrish, D. Carroll, W. B. Karesh, C. Zambrana-Torrelio, W. I. Lipkin, and P. Daszak, "Prediction and prevention of the next pandemic zoonosis," *Lancet*, vol. 380, no. 9857, pp. 1956–1965, 2012, doi: 10.1016/S0140-6736(12)61684-5.
[37] J. Sardar, A. Sau, and P. Mandal, "Clinico-epidemiological profile of confirmed swine flu (H1N1) cases admitted at an infectious disease hospital in Kolkata, India," *Int. J. Community Med. Public Health*, vol. 3, no. 8, pp. 2340–2343, 2016.
[38] M. J. Kane, N. Price, M. Scotch, and P. Rabinowitz, "Comparison of ARIMA and random forest time series models for prediction of avian influenza H5N1 outbreaks," *BMC Bioinf.*, vol. 15, no. 1, p. 276, Dec. 2014.
[39] R. P. Soebiyanto, F. Adimi, and R. K. Kiang, "Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters," *PLoS ONE*, vol. 5, no. 3, p. e9450, Mar. 2010.
[40] *COVID19 India*. Accessed: May 5, 2021. [Online]. Available: https://www.covid19india.org/
[41] W. Wei, J. Jiang, H. Liang, L. Gao, B. Liang, J. Huang, N. Zang, Y. Liao, J. Yu, J. Lai, F. Qin, J. Su, L. Ye, and H. Chen, "Application of a combined model with autoregressive integrated moving average (ARIMA) and generalized regression neural network (GRNN) in forecasting hepatitis incidence in Heng County, China," *PLoS ONE*, vol. 11, no. 6, Jun. 2016, Art. no. e0156768, doi: 10.1371/journal.pone.0156768.
[42] I. N. Soyiri and D. D. Reidpath, "Evolving forecasting classifications and applications in health forecasting," *Int. J. Gen. Med.*, vol. 5, pp. 381–389, 2012, doi: 10.2147/IJGM.S31079.
[43] R. C. Sato, "Disease management with ARIMA model in time series," *Einstein*, vol. 11, no. 1, pp. 128–131, 2013, doi: 10.1590/s1679-45082013000100024.
[44] R. Gupta and S. K. Pal, "Trend analysis and forecasting of COVID-19 outbreak in India," *medRxiv*, Mar. 2020, doi: 10.1101/2020.03.26.20044511.
[45] D. Kwiatkowski, P. C. B. Phillips, P. Schmidt, and Y. Shin, "Testing the null hypothesis of stationarity against the alternative of a unit root," *J. Econometrics*, vol. 54, nos. 1–3, pp. 159–178, Oct. 1992.
[46] G. M. Ljung and G. E. P. Box, "On a measure of lack of fit in time series models," *Biometrika*, vol. 65, no. 2, pp. 67–72, 1978.
[47] R. Adhikari and R. Agrawal, "An introductory study on time series modeling and forecasting," LAP Lambert Acad. Publishing, Saarbrücken, Germany, Tech. Rep. [Online]. Available: https://arxiv.org/abs/1302.6613
[48] D. Ray, M. Salvatore, R. Bhattacharyya, L. Wang, J. Du, S. Mohammed, S. Purkayastha, A. Halder, A. Rix, D. Barker, M. Kleinsasser, Y. Zhou, D. Bose, P. Song, M. Banerjee, V. Baladandayuthapani, P. Ghosh, and B. Mukherjee, "Predictions, role of interventions and effects of a historic national lockdown in India's response to the COVID-19 pandemic: Data science call to arms," *Harvard Data Sci. Rev.*, vol. 2020, 2020, doi: 10.1162/99608f92.60e08ed5.
[49] M. Bhatnagar, "COVID-19: Mathematical modeling and predictions," Apr. 2020, doi: 10.13140/RG.2.2.29541.96488.
[50] V. Papastefanopoulos, P. Linardatos, and S. Kotsiantis, "COVID-19: A comparison of time series methods to forecast percentage of active cases per population," *Appl. Sci.*, vol. 10, no. 11, p. 3880, 2020, doi: 10.3390/app10113880.

[51] Y. Li, B. Wang, R. Peng, C. Zhou, Y. Zhan, and Z. Liu, "Mathematical modeling and epidemic prediction of COVID-19 and its significance to epidemic prevention and control measures," *Ann. Infectious Disease Epidemiol.*, vol. 5, no. 1, p. 1052, 2020.
[52] A. Kotwal, A. K. Yadav, J. Yadav, J. Kotwal, and S. Khune, "Predictive models of COVID-19 in India: A rapid review," *Med. J. Armed Forces India*, vol. 76, no. 4, pp. 377–386, Oct. 2020, doi: 10.1016/j.mjafi.2020.06.001.
[53] R. S. A. Alsudani and J. C. Liu, "The use of some of the information criterion in determining the best model for forecasting of thalassemia cases depending on Iraqi patient data using ARIMA model," *J. Appl. Math. Phys.*, vol. 5, no. 3, pp. 667–679, 2017, doi: 10.4236/jamp.2017.53056.
[54] T. Hu, M. Khishe, M. Mohammadi, G.-R. Parvizi, S. H. T. Karim, and T. A. Rashid, "Real-time COVID-19 diagnosis from X-ray images using deep CNN and extreme learning machines stabilized by chimp optimization algorithm," *Biomed. Signal Process. Control*, vol. 68, Jul. 2021, Art. no. 102764, doi: 10.1016/j.bspc.2021.102764.
[55] C. Wu, M. Khishe, M. Mohammadi, S. H. T. Karim, and T. A. Rashid, "Evolving deep convolutional neutral network by hybrid sine–cosine and extreme learning machine for real-time COVID19 diagnosis from X-ray images," *Soft Comput.*, May 2021, doi: 10.1007/s00500-021-05839-6.
[56] M. Nilashi, S. Samad, L. Shahmoadi, H. Ahmadi, E. Akbari, and T. A. Rashid, "The COVID-19 infection and the immune system: The role of complementary and alternative medicines," *Biomed. Res.*, Vol. 31 no. 3, pp. 1–4, 2020.
[57] T. Rahman, A. Akinbi, M. E. H. Chowdhury, T. A. Rashid, A. Şengür, A. Khandakar, K. R. Islam, and A. M. Ismael, "COV-ECGNET: COVID-19 detection using ECG trace images with deep convolutional neural network," 2021, *arXiv:2106.00436*. [Online]. Available: https://arxiv.org/abs/2106.00436
[58] C. Chakraborty and A. N. Abougreen, "Intelligent Internet of Things and advanced machine learning techniques for COVID-19," *EAI Endorsed Trans. Pervasive Health Technol.*, vol. 7, no. 26, pp. 1–14, 2021, doi: 10.4108/eai.28-1-2021.168505 .
[59] L. J. Muhammad, A. A. Ebrahem, S. U. Sani, A. Abdulkadir, C. Chinmay, and I. A. Mohammed, "Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset," *SN Comput. Sci.*, vol. 2, no. 11, pp. 1–13, 2021, doi: 10.1007/s42979-020-00394-7.
[60] L. Ali, C. Zhu, Z. Zhang, and Y. Liu, "Automated detection of Parkinson's disease based on multiple types of sustained phonations using linear discriminant analysis and genetically optimized neural network," *IEEE J. Transl. Eng. Health Med.*, vol. 7, pp. 1–10, 2019, doi: 10.1109/JTEHM.2019.2940900.
[61] F. S. Ahmad, L. Ali, H. A. Khattak, T. Hameed, I. Wajahat, S. Kadry, and S. A. C. Bukhari, "A hybrid machine learning framework to predict mortality in paralytic ileus patients using electronic health records (EHRs)," *J. Ambient Intell. Hum. Comput.*, vol. 12, no. 3, pp. 3283–3293, Mar. 2021, doi: 10.1007/s12652-020-02456-3.

SUJATA DASH (Member, IEEE) received the Ph.D. degree in computational modeling from Berhampur University, Orissa, India, in 1995. She is currently an Associate Professor with the P.G. Department of Computer Science and Application, North Orissa University, Baripada, India. She has published more than 150 technical articles in international journals, conferences, and book chapters of reputed publications. She has guided many scholars for their Ph.D. degrees in computer science. She is associated with many professional bodies like IEEE, CSI, ISTE, OITS, OMS, IACSIT, IMS, and IAENG. She is in the editorial board of several international journals and also reviewer of many international journals. Her current research interests include machine learning, distributed data mining, bioinformatics, intelligent agent, Web data mining, recommender systems, and image processing.

CHINMAY CHAKRABORTY is currently an Assistant Professor with the Department of Electronics and Communication Engineering, Birla Institute of Technology, Mesra, India. He has published 120 articles at reputed international journals, conferences, book chapters, and books. His current research interests include the Internet of Medical Things, wireless body area networks, wireless networks, telemedicine, m-health/e-health, and medical imaging. He received the Best Session Runner-up Award, the Young Research Excellence Award, the Global Peer Review Award, the Young Faculty Award, and the Outstanding Researcher Award.

SOURAV KUMAR GIRI received the bachelor's degree in computer science and engineering from the National Institute of Technology Rourkela, Odisha, in 2009, India, the M.Tech. degree from the Indian Institute of Technology Kharagpur, in 2015. He is currently a Research Scholar with the Department of Computer Science and Application, North Orissa University, Baripada, India. His current research interests include algorithm analysis, artificial intellegience, machine learning, and blockchain.

SUBHENDU KUMAR PANI received the Ph.D. degree from Utkal University, Odisha, India, in 2013. He is currently working as the Principal with Krupajal Computer Academy (KCA), Bhubaneswar. He has published 51 International Journal articles (25 Scopus index). His professional activities include roles as the Book Series Editor (CRC Press, Apple Academic Press, and Wiley-Scrivener), an associate editor, an editorial board member, and a reviewer of various international journals. He is an associate with number of conference societies. He has more than 150 international publications, five authored books, 15 edited and upcoming books, and 20 book chapters into his account. His research interests include data mining, big data analysis, Web data analytics, fuzzy decision making, and computational intelligence. He is a fellow in SSARS Canada and a Life Member in IE, ISTE, ISCA, OBA.OMS, SMIACSIT, SMUACEE, and CSI. He was a recipient of five researcher awards.

JAROSLAV FRNDA (Senior Member, IEEE) was born in Martin, Slovakia, in 1989. He received the M.Sc. and Ph.D. degrees from the Department of Telecommunications, VSB–Technical University of Ostrava, in 2013 and 2018, respectively. He is currently an Assistant Professor with the University of Žilina, Slovakia. He has authored and coauthored 12 SCI-E and nine ESCI articles in WoS. His research interests include quality of multimedia services in IP networks, data analysis, and machine learning algorithms.