International Journal of E-Health and Medical Communications
Volume 13 • Issue 2 • July-December 2022

Time-Series Forecasting and Analysis of COVID-19 Outbreak in Highly Populated Countries:
A Data-Driven Approach

Arunkumar P. M., Karpagam College of Engineering, Coimbatore, India
Lakshmana Kumar Ramasamy, Hindusthan College of Engineering and Technology, Coimbatore, India
Amala Jayanthi M., Kumaraguru College of Technology, Coimbatore, India

ABSTRACT

A novel corona virus, COVID-19, is spreading across different countries in an alarming proportion, and it has become a major threat to the existence of human community. With more than eight lakh death count within a very short span of seven months, this deadly virus has affected more than 24 million people across 213 countries and territories around the world. Time-series analysis, modeling, and forecasting are important research areas that explore the hidden insights from larger set of time-bound data for arriving at better decisions. In this work, data analysis on COVID-19 dataset is performed by comparing the top six populated countries in the world. The data used for the evaluation is taken for a time period from 22nd January 2020 to 23rd August 2020. A novel time-series forecasting approach based on auto-regressive integrated moving average (ARIMA) model is also proposed. The results will help the researchers from the medical and scientific communities to gauge the trend of the disease spread and improvise containment strategies accordingly.

KEYWORDS

ARIMA, COVID-19, Data Analysis, Disease Spread, Time-Series Forecasting

1. INTRODUCTION

The emergence of novel corona virus is identified from the Wuhan City, Hubei province in China during December 2019 and subsequently renamed as COVID-19 by World health organization. The most common symptoms of the virus include fever, cough and tiredness. Some lesser known symptoms are headache, diarrhea, sore throat and loss of taste or smell. Most of the severe cases of COVID-19 showed symptoms of breathing difficulty and chest pain. Monitoring of epidemiological changes in an in-depth manner will give better perceptions on the disease outbreak (Rotha & Byrareddy, 2020). The research on time-series data is highly critical due to the enormous usage of temporal data in wide variety of applications. Large dataset, high dimensionality and frequent updation are few characteristics of time-series data. The time-series data is subjected to various processing steps to discover the patterns for better decision making. Apart from pattern discovery and clustering, other important task of time-series data mining include classification, rule mining and summarization (Fu,2011). Distance-based clustering, fuzzy c-means (FCM) algorithm, Autoregressive integrated

DOI: 10.4018/IJEHMC.20220701.oa3

moving average (ARIMA) models and Hidden Markov model (HMM) are few methods adopted for time-series clustering and pattern discovery. Time series forecasting depends on the task of analyzing past observations of a random variable and generates a model that portrays the underlying relationship and its patterns. Each of the forecasting method follows four important steps namely, problem definition, information gathering, selecting the best model and forecasting (Hyndman & Athanasopoulos,2018). The time-series analysis and forecasting for COVID-19 disease outbreak is an emerging research paradigm that requires deep knowledge and better experimentations for interpreting the trend and evaluating the predictions.

Holt–Winters Additive Model (HWAAS), Auto-regressive integrated moving average (ARIMA), TBAT, Prophet, DeepAR and N-Beats and Vector Auto regression (VAR) are few models used by researchers around the world for time-series forecasting(Papastefanopoulos,2020). In HWAAS model, trend and seasonal variation of the data are taken in to account. This method is an advanced model proposed by adopting added features to Holt's exponential smoothing. In exponential smoothing, the recently recorded observations are used for updating the prediction levels. The additive method is favored when the seasonal variations are approximately constant through the data series. Holt-Winters Exponential Smoothing is also called as Triple Exponential Smoothing. TBAT method involves four components, namely, Trigonometric seasonal formulation, Box–Cox transformation, ARMA errors and trend component (Harvey et al., 1997; Box & Cox,1964; Adhikari & Agrawal,2013). Multiple seasonalities can be accommodated by TBAT model. Here, each seasonality is modeled with a trigonometric representation based on fourier series. Prophet method is proposed by Facebook. Three major components used by this model are trend, seasonality and holidays.

Time-series decomposition of Prophet model is given by the equation as:

$$Y(t) = A(t) + B(t) + C(T) + \hat{e}\,(t) \tag{1}$$

where, A(t), B(t) and C(T) denotes trend, seasonality and holiday. The last term ê (t) implies error value. Saturating growth model and a piece-wise linear model are utilized by Prophet approach to gather forecasting results. DeepAR deploys a long short-term memory based recurrent neural network architecture for time-series forecasting. Probabilistic forecasting is supported by this model that trains an auto regressive recurrent network (Salinas et al., 2020). N-Beats model is the short form of Neural basis expansion analysis for interpretable time series forecasting. A deep neural architecture consisting of forward and backward residual links is used by N-Beats model (Oreshkin et al.,2019). In this method, generic architecture and an interpretable architecture are used in tandem and dual residual stacking results are observed. Vector Auto regression (VAR) model is a simplification of the univariate autoregressive model for forecasting a vector of time series. It is a multivariate forecasting algorithm .Such model should possess at least two time series variables that influence each other. In ARIMA model, information in the past values of the time series can alone be used to predict the future values.

2. MATERIALS AND METHODS

The COVID-19 data is retrieved from Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. The data contains cumulative count of infected people observed in daily basis. The data is fetched for a time period from 22nd January 2020 to 23rd August 2020. The work is segregated in twofold. The first part of the work demonstrates the data analysis of COVID-19 impact in six highly populated nations. For this work, six months of time-series COVID-19 data (starting from 22-01-2020) is used for evaluation. The second part of the work utilizes ARIMA model for time-series prediction of COVID-19 disease by analyzing two nations, namely, India and US. The COVID-19 dataset consists of daily time series summary of confirmed, recovered and mortality

counts. Province level and country level segregation is also available in the data source. The data is updated once in a day for all the countries listed in the repository. Orange tool is used in this work for exploring and analyzing the data. Orange is an open source toolkit used for explorative data analysis and interactive data visualization (Orange,2020). Python-based scripts are also used during the implementation process for specific functionalities. The data set consists of 267 rows denoting different countries(Hopkins,2020). Few countries in the list have sub-classification based on provinces. The exploration of COVID spread in large inhabitant group requires critical attention. Hence, the top six countries in the world in terms of population are selected for the time-series analysis. As the corona virus got originated from China's Hubei Province, it is selected as the specific zone for the experimentation. Along with Hubei (China), the most populated countries in the world, namely, India, US, Indonesia, Pakistan and Brazil are chosen for the analysis. The experimentation is aimed to track the progress of COVID-19 infection and to visualize the comparative analysis of the disease spread and the impact of control measures.

3. DATA ANALYSIS OF COVID-19 DISEASE IN HIGHLY POPULATED COUNTRIES

The raw dataset is subjected to required pre-processing steps. The given data table is reinterpreted as time series object. Time-series analysis is performed for COVID-19 data with respect to the most populated countries. Initially, the plot of COVID-19 confirmed cases(y axis) with the time period in days (x axis) for the six countries is visualized. The graph is aimed to reflect the disease spread among the top six populated countries. The disease started affecting each country at different time period. Hence, the y axis component is log transformed for better interpretation and visualization. Figure 1 portrays the spreading of virus among the selected countries for the given time period.

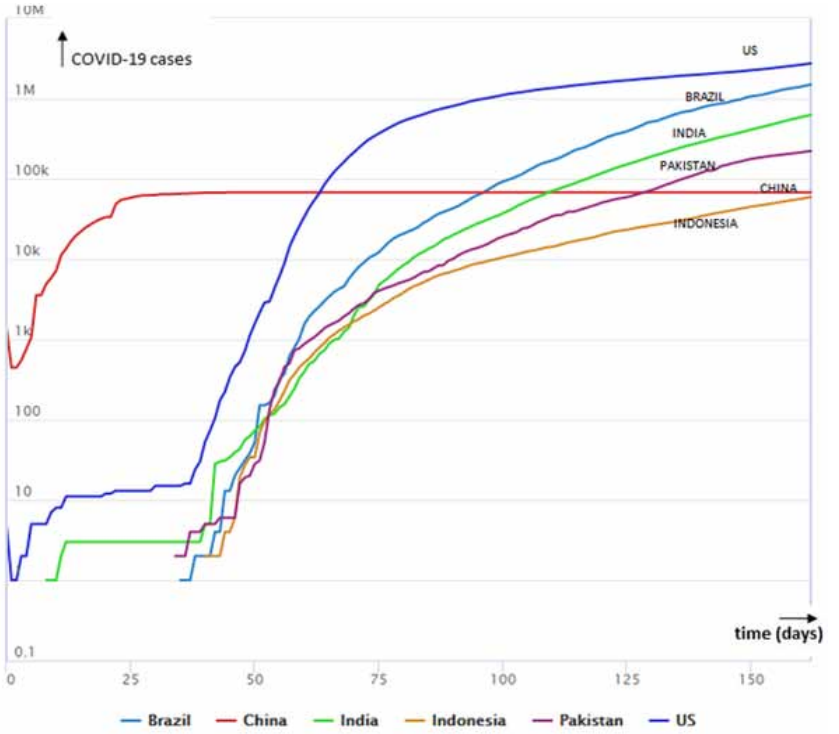
Hubei province in China reported 444 active COVID-19 cases on the first day of evaluation (22nd January 2020). The virus was spreading in astonishing fashion since its outbreak in China. As the above figure suggests, except China, the initial period between 25th day and 50th day shows the slow growth of disease spread among the highly populated countries. During the 25th day of evaluation, China reported a staggering COVID-19 count of 58182 in Hubei province. US and India reported 13 and 3 active cases respectively on the same day. After the enormous peaking in the pandemic statistic count, China (Hubei) reported stabilization and reached considerable slowness in the disease spread and the stableness continued for a long period. The active cumulative count value of China (Hubei) stands 68135 on 2nd July 2020. But for all other countries, the progression is more severe during the 50th to 75th day timeslot. On the 140th day of the observed time period, US crossed 2 million active cases as the collective tally. Table 1 portrays the cumulative count of COVID-19 among the most populated countries.

Also, US reported incredible increase in COVID-19 cases with a cumulative count of 2742049 (more than 2.7 million) on 2nd July 2020.On the same day, India has recorded 625544 cases (above 6 lakhs) as the cumulative count. Also, the statistics reflects the emergence of Brazil in battling the deadly virus among the most populated countries. Brazil is the only country after US to possess more than 1 million COVID-19 cases as on 2-07-2020.Pakistan also shows gradual increase in the count, whereas Indonesia is better placed with less infected count compared to other high-populated nations. As a matter of fact, the date of first infection of COVID-19 will differ among the chosen countries. Hence, by using python code, the time-series analysis is carried out by applying a threshold of (k=100) for the above dataset, where 'k' denotes the active cases reported by individual countries. This will allow better visualization of the data with the graph showing the growth of COVID-19 cases starting from the same point of reference. The modified result is shown in Figure 2.

3.1. Impact of Smoothing and Differencing in Time-Series COVID-19 Data

The preprocessing steps are needed to process the COVID-19 data effectively. The time series data is made as stationary component. First or second order discrete differences are captured for this task.

Figure 1. COVID-19 spread among most-populated countries



The critical changes in the output of the graph can be analyzed with more clarity by this option. In this work, order of differencing is chosen as '1' and with the discrete differencing, the curves are plotted for analysis. The prominent spikes and abnormal values can be identified by comparing the original and differential versions. The comparison of China(Hubie) and its transformed version is shown in Figure 3.

The above figure shows the comparison of the curves (differential values) with normal values while considering a specific country, say, China. The abnormal spike in the differenced time-series output denotes the unusual changes in day-to-day statistics reported from China. The similar comparison is performed with respect to Indian COVID-19 data as shown in Figure 4. Here, the curve output is

Table 1. Timeline of COVID-19

Country	Cumulative COVID-19 confirmed cases (starting from 22-01-2020)			
	50th day count	75th day count	100th day count	150th day count
Brazil	52	12161	92202	1067579
China (Hubei)	67781	67803	68128	68135
India	73	4778	37257	410451
Indonesia	34	2491	10551	45029
Pakistan	28	4035	19103	176617
US	1561	367215	1106829	2255327

Figure 2. Modified chart of disease spread (at least 100 cases)

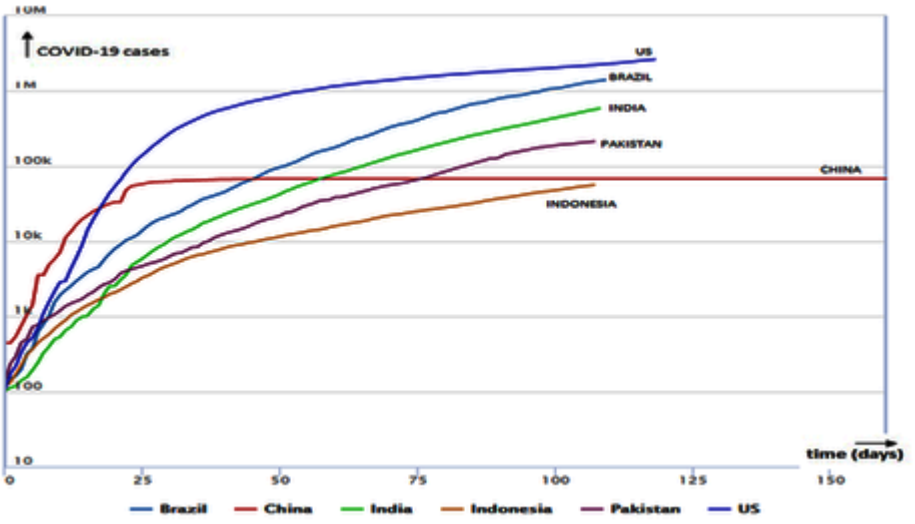
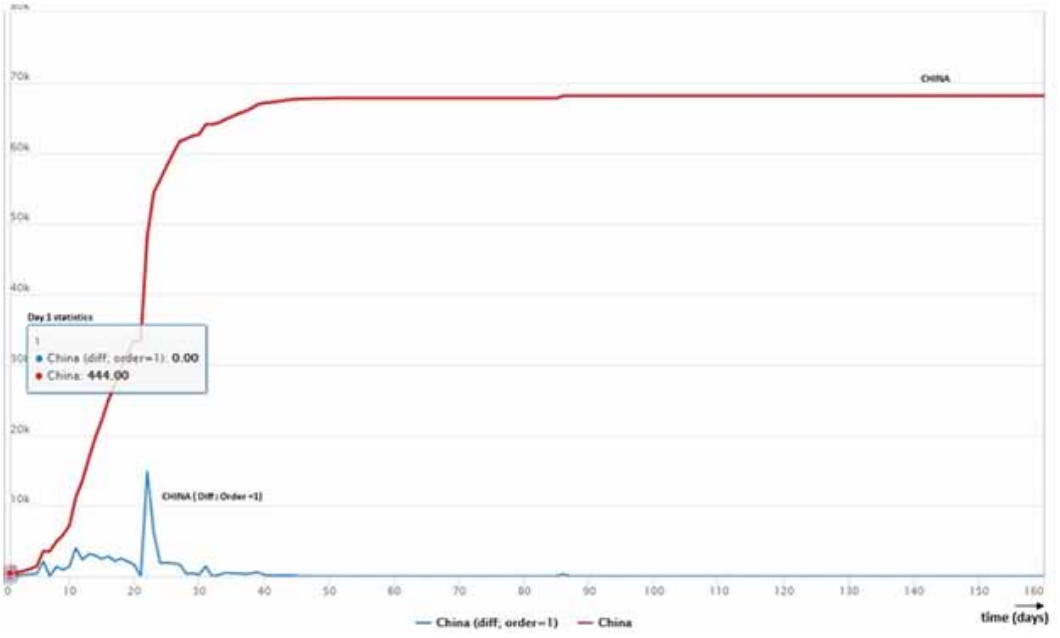


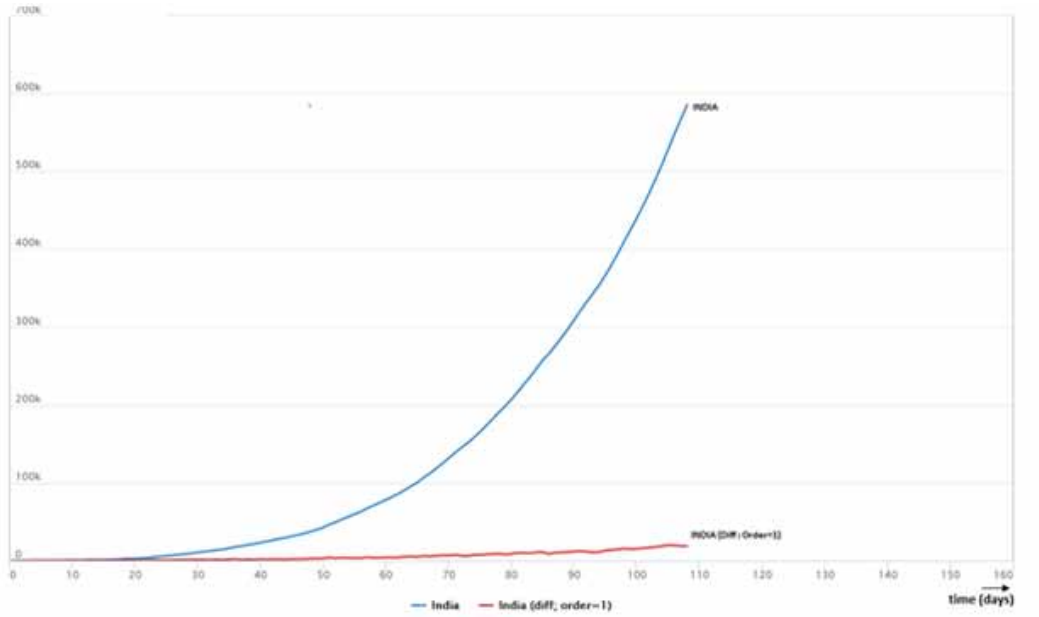
Figure 3. Comparison of original and differential analysis (China)



smooth and it reflects the undisturbed growth of cases each day. The inconsistency is hardly seen while analyzing the Indian COVID-19 data using differenced time series output.

The time series analysis is performed by comparing the stationary time-series output (differenced data) between different countries. This will show the way the disease is spreading and how the disease

Figure 4. Comparison of original and differential analysis(India)



is controlled by respective countries. The comparison between India, US and China is shown in Figure 5. Similarly, Figure 6 portrays the comparison of Brazil, Indonesia and Pakistan.

The results show that, more fluctuations are observed during the analysis pertaining to US and Brazil. The inconsistency and unpredictability are more in these countries while analyzing the day-to-day statistics. Also, China and Pakistan are the countries with odd spikes present at some time interval. Though the disease count is increasing in India, the time series analysis show better control with minimum 'rise and fall' within the given time slice. The result also underlines the consistency of COVID-19 data projected and analyzed for the Indian population compared with other largely populated countries. The percentage of change is also utilized in this work to observe the relative growth of the disease. The output of such comparison for China, India and US is shown in Figure 7.

The figure shows that the trend analysis for the three countries by showing the percentage of change during the pandemic period. The poor performance of China at the initial phase is observed followed by better control in later days. The analysis of data from India demonstrated minimum changes for the complete time period during the comparative analysis. Also, the result conveys that the countries like Pakistan and Brazil are not up to the mark in this regard. Also, the above output is further improvised with smoothing effect. This will allow to visualize the change in trend among the countries more accurately. 'Moving transform' property is used which applies rolling window functions to the time series. Moving average function with a mean of every 5 values in the series is selected and output is observed. After applying smoothing, the differential output with percentage change is recorded and analyzed. This visualization offers better perception in viewing the changes in the trend. The output is depicted in Figure 8.

The data visualization of the countries after applying the moving transform function is recorded in the previous step. The difference in percentage is observed by comparing the countries' COVID-19 count directly in relative manner. Also, the same comparison (after smoothing) is carried out by applying differential function to record the absolute growth instead of seeing the output as change in percentage. The output is shown in Figure 9.

Figure 5. Comparison of differential analysis (China, India and US)

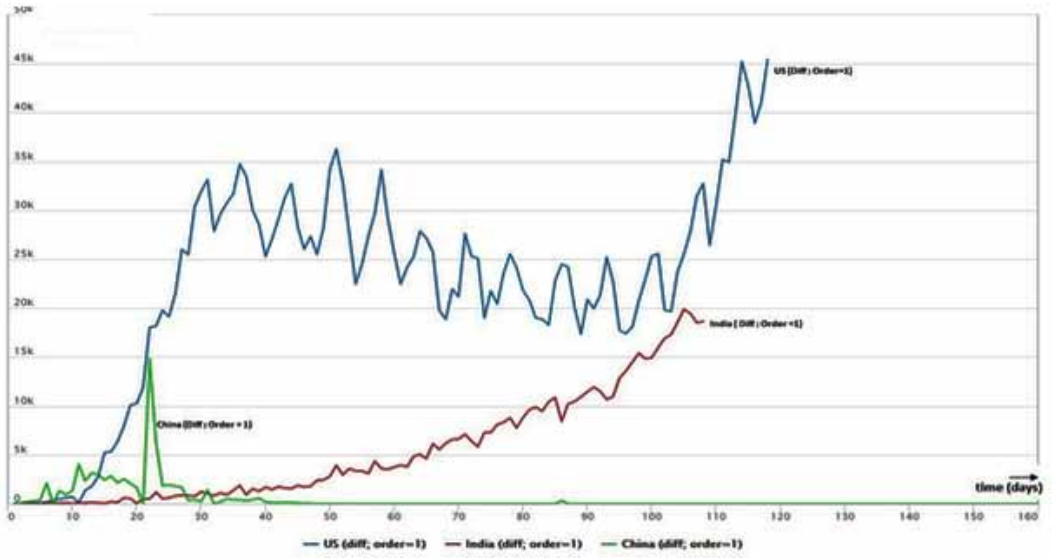


Figure 6. Comparison of differential analysis (Brazil, Indonesia and Pakistan)

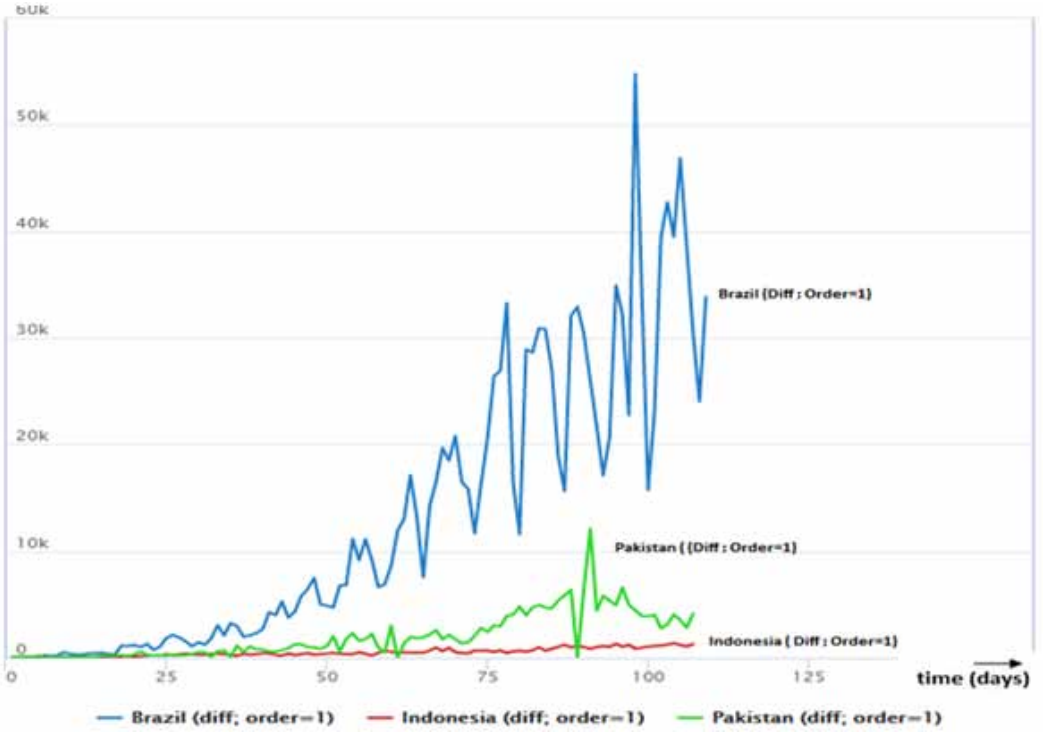


Figure 7. Percentage of change in COVID-19 outbreak

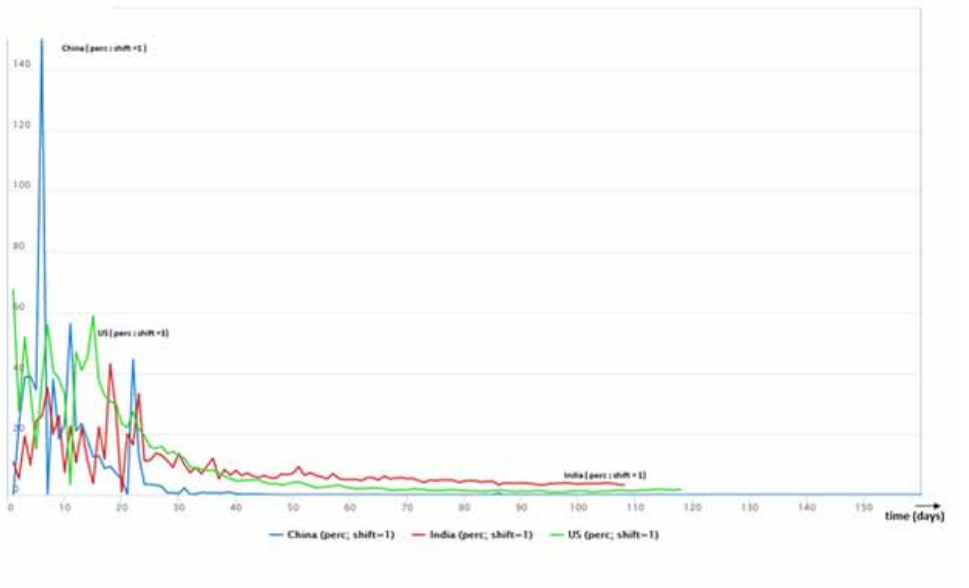


Figure 8. Effect of smoothing in time-series COVID-19 data

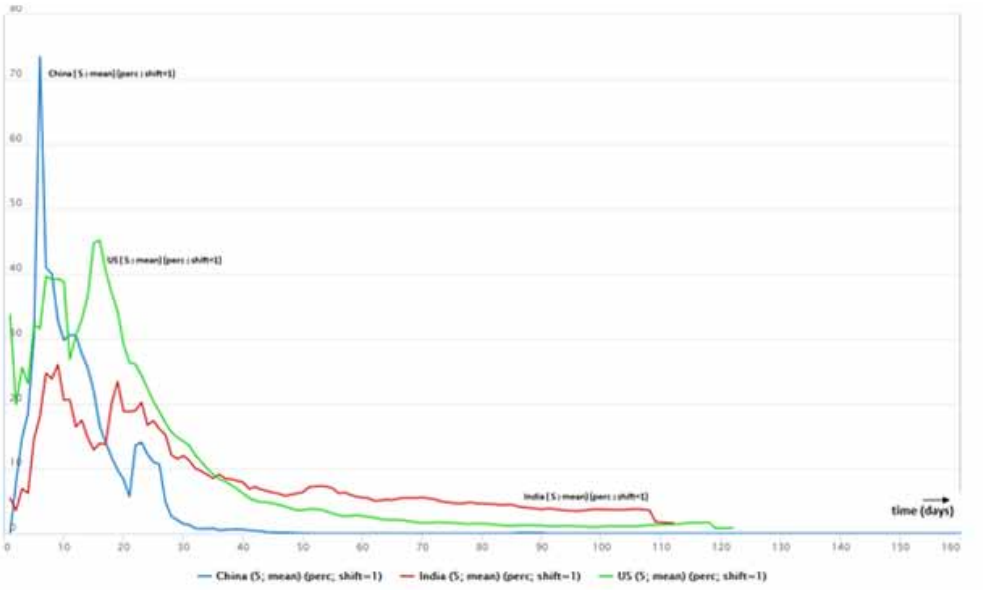
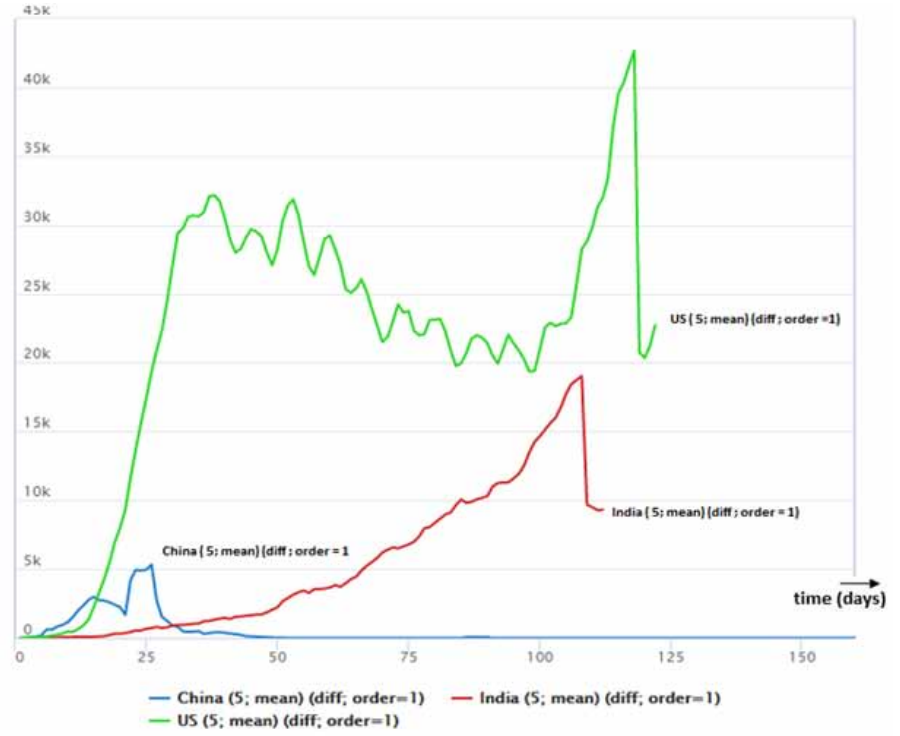


Figure 9. Absolute growth chart after smoothing



The above chart clearly shows the change in trend for each of the country selected for observation. Here, on day 16, US takes a giant leap and overtakes the progression of Chinese curve. Similarly on day 30, India crosses the count of China and makes a clear increment thereafter. The time series analysis of this nature helps the scientists and researchers to understand the trend and offer better insights for further evaluations.

4. TIME-SERIES FORECASTING OF COVID-19 OUTBREAK

Telehealth and Telemedicine play a vital role in providing remote health services to the patients by using modern technologies. The risk of contagious infections can be mitigated by deploying telemedicine technologies. During this pandemic situation, the remote assessment of patients who are vulnerable to infections is the need of the hour. Integration of telehealth with the existing health care system is needed for fruitful results. In the past few years, video consultations and chatbot are used for remote consultations. General unwillingness and financial constraints are the strong barriers in implementing telehealth systems. Even the western countries are finding hard in realizing the full-fledged telemedicine systems. Proper education and training are required for better deployment of telehealth system during the spreading of infectious diseases. Improving the funding facility and managing the needs of all stakeholders are the immediate solutions (Ahmed et al., 2010).

In recent years, Machine learning techniques have evolved as a very good alternative to conventional statistical approaches in the area of forecasting and analysis. Comparison of various machine learning algorithms for time series forecasting is performed using M3 competition data. Radial basis functions, kernel regression, Bayesian neural networks, CART, KNN and Gaussian processes are deployed for the study (Cerqueira et al., 2019). Preprocessing steps such as Log

transformation, Deseasonalization and Scaling are implemented during the performance evaluation. MLP and GP show the best performance during the execution process. The amount of sample size is critical in choosing the analyzing technique for time series forecasting. For large data size, machine learning methods are more suited than statistical models (Bontempi et al., 2012). An empirical study is conducted to gauge the performance of statistical and pure ML models with respect to sample size. The deliberation of the analysis reflects the importance of sample size in univariate time series forecasting. Auto-Regressive Integrated Moving Average model, Naive2, Theta, exponential smoothing state-space model and Tbats are the models used under statistical analysis. Multivariate adaptive regression splines, Gaussian Process regression, Random forest, Generalized linear model and rule-based model are deployed as pure ML algorithms. More than 1000 observations from tsdl database are taken for the experimentation. The comparative analysis of individual model and multi-step-ahead forecasting is carried out. The results underline the importance of data size for time series forecasting. Machine learning models can be used for forecasting applications by considering three important aspects. Assuming one-step forecasting as supervised learning process is the foremast facet. Instantiation with Local Learning and smooth transition from one-step to multiple-step forecasting are the other characteristics to be considered (Smith et al., 2020).

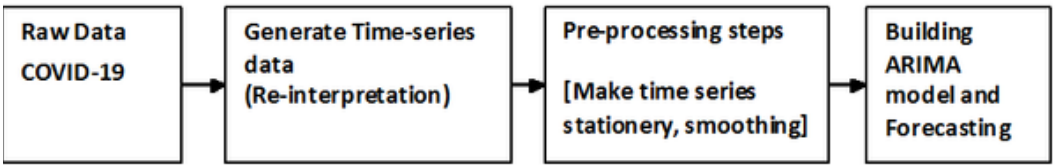
In the past few months, modeling and forecasting of COVID-19 using advanced techniques was carried out by many researchers. A novel autoregressive time series models based on the two–piece scale mixture normal (TP–SMN) distributions is proposed (Maleki et al., 2020). Mean relative percentage error (MAPE) is used to assess the performance of the model. Prediction results show MAPE value 0.22% for the assessment of confirmed COVID-19 cases, which is reasonably good in terms of performance. Time series forecasting of COVID-19 using deep learning method is reported (Chimmula & Zhang,2020). Long short-term memory (LSTM) network is used in this work to evaluate the disease forecasting in Canada. RMSE error of 45.70 with an accuracy of 92.67% is inferred at the end of the experimentation for long term prediction. Genetic programming based model is developed to forecast the corona virus spread in India (Salgotra et al., 2020). The states of Maharashtra, Gujarat and Delhi are taken for evaluation. The usage of linkage function underlines the superiority of genetic programming approach for time-series prediction. A novel exponential smoothing model for COVID-19 disease forecasting is performed (Petropoulos & Makridakis,2020). Five rounds of forecasting is carried out until March 2020 and the results are reliable to initiate further deliberations. A Comparative study on five different deep learning algorithms for the purpose of COVID-19 forecasting is done. Recurrent Neural Network (RNN), Long short-term memory (LSTM), Bidirectional LSTM (BiLSTM), Gated recurrent units (GRUs) and Variational AutoEncoder (VAE) are used in the study (Zeroual et al.,2020). Italy, Spain, France, China, USA, and Australia are the countries selected for evaluation. VAE model outperformed other methods during the assessment.

The combination of ARIMA model and Wavelet-based forecasting model is proposed to address non-linearity and non-stationary actions of univariate data (Chakraborty & Ghosh2020). A modified stacked auto-encoder based forecasting model is deployed to assess the COVID-19 data pertaining to China (Hu et al., 2020). Machine learning based prediction model for COVID-19 is implemented using standard algorithms such as logistic regression, decision tree, support vector machine, naive Bayes, and artificial neutral network (Muhammad et al., 2020). The data is fetched from the repository of epidemiology pertaining to Mexico. Correlation coefficient analysis is carried out to capture the relationship between different features. The performance of the model is evaluated using metrics such as accuracy, sensitivity and specificity. The result highlighted the superior performance of decision tree model with 95% accuracy. The results are highly encouraging for Artificial Intelligence research community to extend the research findings in Time-series forecasting and modeling.

4.1 Proposed Time-Series Forecasting Approach Using ARIMA Model

Time Series forecasting is the method of deploying statistical model to predict future values of a time-series data based on past results. Time-series forecasting is divided in to two types namely, Univariate

Figure 10. Proposed ARIMA model for COVID-19 forecast



and multivariate forecasting. There are many methods available for time series forecasting. Few prominent techniques used are Exponential Smoothing, Single Equation Regression, Simultaneous-equation Regression, Autoregressive Integrated Moving Average (ARIMA) and Vector Autoregression (VAR).

In this research work, ARIMA model is utilized to forecast time series data pertaining to COVID-19. This model is one of the prominent forecasting methods used in time-series analysis. The features of ARIMA model are the ability to support Box–Jenkins methodology and better usage of exponential smoothing. This model exploits linear correlation between the time-series values and utilizes the linear dependencies in observations for extracting better results. Efficient model selection and high level of interpretability are the major advantages of ARIMA model. China, India and US are the three most populated countries in the world. The statistics of China taken during the last two months (July andAugust-2020) implies that the virus spread is under control and it is in a state of saturation. The other two large countries, namely, India and US are reporting high amount of COVID-19 positive cases. Hence, for the modeling and disease forecasting, both India and US are considered for evaluation.

The general steps of the proposed time series forecasting model is shown in Figure 10.

Initially, the raw COVID-19 data is obtained from Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. The data is analyzed for a period from 22-01-2020 to 23-08-2020.Time-series forecasting is performed for the next 20 days (24-8-2020 to 12-09-2020). The data is reinterpreted as time series component and the complete COVID-19 data is revamped as time-series data. After data aggregation, pre-processing steps are initiated. Making time-series data as stationary by differential data calculations and smoothing is essential. This work is initiated by applying rolling window functions. Forecasting is instigated with the help of ARIMA model.

The ARIMA model is depicted as ARIMA(p,d,q),where Here, ‘p’ denotes order of auto-regression, ‘d’ shows the degree of trend difference and ‘q’ is the order of moving average.AR denotes auto regression process and it is the lag order that utilizes the dependent relationship between observations. ‘ I’denotes integrated factor which uses differencing of original data to make the time series flow as stationary. ‘MA’stands for moving average and it exploits the dependency of a particular data with the residual error from the model applied to lagged points.

In an Auto regressive model, the forecasts correspond to a linear combination of past values of the variable. In a Moving Average model the forecasts correspond to a linear combination of past forecast errors. The time series is seen as stationary by using differencing (Integrating) step.

An autoregressive model of order p is given as:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \quad (2)$$

ε_t is white noise, parameters $\phi 1,\ldots,\phi_p$ are varied to get different time series patterns, regression is formulated with lagged values of y_t as predictors.

In AR(p) model, y_t depends only on its own lags.

In a moving average model, past forecast errors in a regression-like model is utilized:

Table 2. Selection of ARIMA parameters (country:India)

COVID-19 data: INDIA (22-01-2020 to 23-08-2020)	
ARIMA parameters (p,d,q)	**Mean Squared Error values**
ARIMA(0, 0, 0)	MSE=1811554255481.896
ARIMA(0, 0, 1)	MSE=465185544947.711
ARIMA(0, 0, 2)	MSE=126402516577.023
ARIMA(0, 1, 0)	MSE=1264462774.128
ARIMA(0, 1, 1)	MSE=323001217.980
ARIMA(0, 2, 0)	MSE=11547117.349
ARIMA(0, 2, 1)	MSE=12600663.200
ARIMA(0, 2, 2)	MSE=10931400.989
ARIMA(0, 2, 3)	MSE=11312028.921
ARIMA(1, 0, 0)	MSE=1908253749.919
ARIMA(1, 1, 0)	MSE=11789300.000
ARIMA(1, 2, 0)	MSE=12256255.397
ARIMA(2, 1, 0)	MSE=12495106.773
ARIMA(2, 2, 0)	MSE=11714854.922
ARIMA(2, 2, 1)	MSE=13274147.571
ARIMA(4, 2, 0)	MSE=10870436.536
ARIMA(6, 1, 0)	MSE=12470698.418
Best model is (4, 2, 0) with minimum MSE	

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \quad (3)$$

Hence, in MA(q) model, y_t is weighted moving average of the past few forecast errors. So, y_t depends only on the lagged forecast errors.

By including differencing operation with auto regression and a moving average model, ARIMA model is developed. ARIMA model equation is given as:

$$y't = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (4)$$

$y't$ denotes differenced series, lagged values of y_t and lagged errors are shown as predictors.

The equation of ARIMA model can be simplified by the following relation (Selva, 2020):

Predicted Yt = Constant + Linear combination Lags of Y (upto 'p' lags) + Linear Combination of Lagged forecast errors (upto 'q' lags).

So, in ARIMA model, time series was differenced at least once to make it stationary and amalgamation of AR and MA term is performed. Hence, in this model, 'p' denotes order of autoregressive part, 'd'denotes degree of first differencing and 'q'denotes order of the moving average part.

Table 3. Forecast of COVID-19 outbreak (India)

Date	India (forecast)	India (95% CI low)	India (95% CI high)
24-08-2020	3168880	3165390	3172380
25-08-2020	3234730	3226890	3242560
26-08-2020	3300640	3288620	3312670
27-08-2020	3369090	3352540	3385630
28-08-2020	3437760	3416900	3458630
29-08-2020	3504620	3479520	3529720
30-08-2020	3572010	3542180	3601840
31-08-2020	3639680	3604700	3674660
1-09-2020	3707520	3667040	3748000
2-09-2020	3776600	3730230	3822970
3-09-2020	3845980	3793560	3898400
4-09-2020	3915320	3856690	3973950
5-09-2020	3985090	3920010	4050170
6-09-2020	4054950	3983210	4126680
7-09-2020	4125070	4046440	4203700
8-09-2020	4195770	4110000	4281530
9-09-2020	4266750	4173660	4359850
10-09-2020	4338060	4237440	4438670
11-09-2020	4409720	4301400	4518050
12-09-2020	4481610	4365400	4597830

In this work, an efficient method is deployed to automate the process of training and evaluating ARIMA models on different combinations of model hyper parameters. This process is also termed as grid search or model tuning. Iteration of ARIMA parameters is carried out using Python implementation and the results are tabulated. The configuration that gives the least Mean Squared Error(MSE) values is the best model and it is selected for capturing the forecast results(Brownlee,2017). The selection of best combination of p,d,q values is executed. The selection of best configuration of ARIMA model for forecasting COVID-19 disease spread in India is shown in Table 2.

The best values of p,d,q determined by the above process is taken from the table and forecasting process is initiated for COVID-19 prediction in India. The overall data taken for the evaluation is from 22-01-2020 to 23-08-2020.The forecast analysis is done for the next 20 days (i.e) from 24-08-2020 to 12-09-2020 with 95% confidence interval. The forecasting results (India) are tabulated in Table 3.

The same procedure is applied for the data from US and the time-series analysis and forecasting is done. The results are tabulated in Table 4 and Table 5.

Forecast results for US data from 24-08-2020 to 12-09-2020 with 95% confidence interval are listed and tabulated as shown in Table 5.

The above results of forecasting model using ARIMA underlines the proportion of COVID-19 disease outbreak that can be expected in coming days. The disease spread in the two most populated countries, namely, India and US is analyzed and the predicted outbreak for the next 20 days is observed and recorded. The magnitude of rise in disease count is much higher for US when compared to India until July 2020 and during the month of August 2020, India reports more cases per day worldwide

than any other countries. Moreover, the strategies for controlling the COVID-19 differs from one country to other. Though both US and India are still in the escalating stage with respect to disease spread, stringent measures will help in controlling the dispersion of the disease.

Table 4. Selection of ARIMA parameters (country:US)

COVID-19 data: US (22-01-2020 to 23-08-2020)	
ARIMA parameters (p,d,q)	**Mean Squared Error values**
ARIMA(0, 0, 0)	MSE=7869189786148.979
ARIMA(0, 0, 1)	MSE=1991482886978.243
ARIMA(0, 1, 0)	MSE=1052908881.723
ARIMA(0, 1, 1)	MSE=370544437.875
ARIMA(0, 2, 0)	MSE=37526913.164
ARIMA(0, 2, 1)	MSE=38438541.222
ARIMA(0, 2, 2)	MSE=40953759.065
ARIMA(0, 2, 3)	MSE=35305186.200
ARIMA(1, 0, 0)	MSE=2720312543.069
ARIMA(1, 1, 0)	MSE=37276121.098
ARIMA(1, 2, 0)	MSE=38386596.142
ARIMA(1, 2, 1)	MSE=36513044.125
ARIMA(1, 2, 2)	MSE=36994984.095
ARIMA(1, 2, 3)	MSE=37360005.969
ARIMA(2, 1, 0)	MSE=38412162.020
ARIMA(2, 2, 0)	MSE=39301704.596
ARIMA(2, 2, 1)	MSE=37014283.917
ARIMA(4, 1, 0)	MSE=35280765.742
ARIMA(4, 2, 0)	MSE=33378921.642
ARIMA(4, 2, 1)	MSE=32654828.334
ARIMA(6, 1, 0)	MSE=30843365.086
ARIMA(6, 2, 0)	MSE=30991965.714
ARIMA(6, 2, 1)	MSE=25104161.267
ARIMA(8, 1, 0)	MSE=24645139.363
ARIMA(8, 2, 0)	MSE=20139905.150
ARIMA(10, 1, 0)	MSE=20076757.494
Best model is (10, 1, 0) with minimum MSE	

Table 5. Forecast of COVID-19 outbreak (US)

Date	US (forecast)	US(95%CI low)	US(95%CI high)
24-08-2020	5732360	5726990	5737720
25-08-2020	5768200	5758480	5777910
26-08-2020	5807550	5792640	5822460
27-08-2020	5847730	5827370	5868080
28-08-2020	5888000	5861590	5914420
29-08-2020	5925220	5892650	5957790
30-08-2020	5955110	5915280	5994950
31-08-2020	5981310	5932310	6030310
1-09-2020	6010490	5950570	6070410
2-09-2020	6043920	5972200	6115640
3-09-2020	6079000	5994880	6163130
4-09-2020	6113560	6016540	6210590
5-09-2020	6144790	6034370	6255200
6-09-2020	6170450	6045730	6295170
7-09-2020	6192940	6052450	6333440
8-09-2020	6217410	6059680	6375140
9-09-2020	6245950	6070020	6421880
10-09-2020	6276740	6082020	6471460
11-09-2020	6306900	6092940	6520850
12-09-2020	6333770	6100090	6567450

5. CONCLUSION

In this work, time series modeling and forecasting of COVID-19 outbreak is carried out. The time-series analysis of top six populated countries is performed and the outcome is analyzed for predicting the trend. The impact of differencing and smoothing in a time-series model is highlighted. A novel ARIMA based time-series forecasting is developed. Using grid search, the large number of hyper parameters are evaluated with the correlation of Mean squared error. The results underline the importance of time-series modeling for predicting the outcome of pandemic outbreak. The deliberations of this work will assist the government authorities to put forward better policy and procedures for containing the infectious COVID-19 disease. In future, the time-series forecasting can be tested using advanced Deep learning algorithms. The deep learning approach such as LSTM can be utilized effectively in near future for time-series analysis of COVID-19 data.

REFERENCES

Adhikari, R., & Agrawal, R. K. (2013). *An introductory study on time series modeling and forecasting.* arXiv preprint arXiv:1302.6613.

Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews*, *29*(5-6), 594–621. doi:10.1080/07474938.2010.481556

Bontempi, G., Taieb, S. B., & Le Borgne, Y. A. (2012, July). Machine learning strategies for time series forecasting. In *European business intelligence summer school* (pp. 62–77). Springer.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B. Methodological*, *26*(2), 211–243. doi:10.1111/j.2517-6161.1964.tb00553.x

Brownlee, J. (2017). *Introduction to time series forecasting with python: how to prepare data and develop models to predict the future*. Machine Learning Mastery.

Cerqueira, V., Torgo, L., & Soares, C. (2019). *Machine learning vs statistical methods for time series forecasting: Size matters.* arXiv preprint arXiv:1909.13316.

Chakraborty, T., & Ghosh, I. (2020). Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. *Chaos, Solitons, and Fractals*, *135*, 109850. doi:10.1016/j.chaos.2020.109850 PMID:32355424

Chimmula, V. K. R., & Zhang, L. (2020). Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, Solitons, and Fractals*, *135*, 109864. doi:10.1016/j.chaos.2020.109864 PMID:32390691

Fu, T. C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, *24*(1), 164–181. doi:10.1016/j.engappai.2010.09.007

Harvey, A., Koopman, S. J., & Riani, M. (1997). The modeling and seasonal adjustment of weekly observations. *Journal of Business & Economic Statistics*, *15*(3), 354–368.

Hu, Z., Ge, Q., Jin, L., & Xiong, M. (2020). *Artificial intelligence forecasting of covid-19 in china.* arXiv preprint arXiv:2002.07112.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Johns Hopkins University Center for Systems Science and Engineering. (n.d.). *Coronavirus (COVID-19) confirmed Cases*. https://github.com/CSSEGISandData/COVID-19

Maleki, M., Mahmoudi, M. R., Wraith, D., & Pho, K. H. (2020). Time series modelling to forecast the confirmed and recovered cases of COVID-19. *Travel Medicine and Infectious Disease*, *37*, 101742. doi:10.1016/j.tmaid.2020.101742

Muhammad, L. J., Algehyne, E. A., Usman, S. S., Ahmad, A., Chakraborty, C., & Mohammed, I. A. (2020). Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset. *SN Computer Science*, *2*(1), 1-13.

Orange Tool. (n.d.). https://orange.biolab.si/

Oreshkin, B. N., Carpov, D., Chapados, N., & Bengio, Y. (2019). *N-BEATS: Neural basis expansion analysis for interpretable time series forecasting.* arXiv preprint arXiv:1905.10437.

Papastefanopoulos, V., Linardatos, P., & Kotsiantis, S. (2020). COVID-19: A Comparison of Time Series Methods to Forecast Percentage of Active Cases per Population. *Applied Sciences (Basel, Switzerland)*, *10*(11), 3880. doi:10.3390/app10113880

Petropoulos, F., & Makridakis, S. (2020). Forecasting the novel coronavirus COVID-19. *PLoS One*, *15*(3), e0231236. doi:10.1371/journal.pone.0231236 PMID:32231392

Rothan, H. A., & Byrareddy, S. N. (2020). The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. *Journal of Autoimmunity*, *109*, 102433. doi:10.1016/j.jaut.2020.102433 PMID:32113704

Salgotra, R., Gandomi, M., & Gandomi, A. H. (2020). Time Series Analysis and Forecast of the COVID-19 Pandemic in India using Genetic Programming. *Chaos, Solitons, and Fractals*, *138*, 109945. doi:10.1016/j.chaos.2020.109945 PMID:32508399

Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, *36*(3), 1181–1191. doi:10.1016/j.ijforecast.2019.07.001

Selva, P. (2020). *Machine Learning Plus*. https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/

Smith, A. C., Thomas, E., Snoswell, C. L., Haydon, H., Mehrotra, A., Clemensen, J., & Caffery, L. J. (2020). Telehealth for global emergencies: Implications for coronavirus disease 2019 (COVID-19). *Journal of Telemedicine and Telecare*.

Zeroual, A., Harrou, F., Dairi, A., & Sun, Y. (2020). Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study. *Chaos, Solitons & Fractals, 140*, 110121.

Arunkumar P. M. (PhD) is currently working as Associate Professor in the Department of Computer Science and Engineering at Karpagam College of Engineering, Coimbatore, Tamil Nadu, India. He completed his PhD in Information and Communication Engineering at the Anna University, Chennai. He is an accomplished professional from the teaching fraternity with 15 years of experience. He obtained his Master's in Computer Science and Engineering in 2005 and received his BE in Electronics and Communication Engineering in 2003. He carried out his research in the area of wireless video QoE and has published various articles in reputed international journals. His research interests include multimedia QoE, wireless networks, machine learning, algorithm design and data analytics.

Lakshmana Kumar Ramasamy (PhD) is currently working as the Head Centre of Excellence for Artificial Intelligence and Machine Learning in Hindusthan College of Engineering and Technology, Coimbatore. Tamil Nadu. He is working as director – Research and Development (AI) in a Canadian based startup company in the Vancouver region of British Columbia, Canada. He is a global chapter Lead for MLCS [Machine Learning for Cyber Security]. He himself involves in research and expertise in AI and Blockchain technologies. He serves as guest editor for many Q1 and Q2 journals. He is an expert editorial advisory board member of Artificial Intelligence in Cambridge Scholars Publishing, UK. He is the Founding Member- IEEE SIG of Big Data for Cyber Security and Privacy, IEEE. He serves in the editorial board of Trends in Renewable Energy Journal, USA. He holds edited books from CRC, Elsevier and Scrivener Publishing. He holds around 20 patents. He is a Data Science certified from John Hopkins University, United States. He also holds the Amazon Cloud Architect certification from Amazon Web Services. He is an ACM Distinguished speaker and IEE Member.

Amala Jayanthi M. is working as an Assistant professor at Kumaraguru College of Technology, India. She has published papers at national and international levels. Her area of expertise is data mining and especially she works in the field of education and healthcare.