Implementation of Stacking Based ARIMA Model for Prediction of Covid-19 Cases in India

Aman Swaraj (182211001@nitdelhi.ac.in)
National Institute of Technology, Delhi

Arshpreet Kaur
National Institute of Technology, Delhi

Karan Verma
National Institute of Technology, Delhi

Ghanshyam Singh
Malaviya National Institute of Technology Jaipur2

Ashok Kumar
Government Mahila Engineering College, Ajmer

Leandro Melo de Sales
Universidade Federal De Alagoas-UFAL, Brasil

Research Article

Keywords: Hybrid Model, Forecasting, COVID-19, ARIMA, NAR

Posted Date: August 4th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-52063/v1

Version of Record: A version of this preprint was published at Journal of Biomedical Informatics on August 1st, 2021. See the published version at https://doi.org/10.1016/j.jbi.2021.103887.

Implementation of Stacking Based ARIMA Model for Prediction of Covid-19 Cases in India

Aman Swaraj[1*], Karan Verma[1], Arshpreet Kaur[1], Ghanshyam Singh[2], Ashok Kumar[3]
Leandro Melo de Sales[4]
182211001@nitdelhi.ac.in[1*],karanverma@nitdelhi.ac.in[1],
arshpreet@nitdelhi.ac.in[1],gsingh.ece@mnit.ac.in[2], kumarashoksaini@gmail.com[3],
leandro@ic.ufal.br[4]
National Institute of Technology, Delhi[1], Malaviya National Institute of Technology Jaipur[2]
Government Mahila Engineering College, Ajmer[3], Universidade Federal De Alagoas-UFAL, Brasil[4]

***Corresponding Author:**

Aman Swaraj
Masters in Technology,
National Institute of Technology, Delhi

Institute Address:
A-7, Institutional Area,
near Satyawadi Raja Harish Chandra Hospital,
New Delhi, Delhi 110040
India

Permanent Address:
S2-286, Sector-4
Unique Society, Vaishali,
Ghaziabad, Uttar Pradesh – 201010
India

Tel: 91-7217842795, 91-7783860847
Email: 182211001@nitdelhi.ac.in; amanswaraj007@gmail.com

Conflict of Interest

The authors declare that they have no conflict of interest.

ABSTRACT

Background: Time-series forecasting has a critical role during pandemics as it provides essential information that can lead to abstaining from the spread of the disease. The novel coronavirus disease, COVID-19, is spreading rapidly all over the world. The countries with dense populations, in particular, such as India, await imminent risk in tackling the epidemic. Different forecasting models are being used to predict future cases of COVID-19. The predicament for most of them is that they are not able to capture both the linear and nonlinear features of the data solely.

Methods: We propose an ensemble model integrating an autoregressive integrated moving average model (ARIMA) and a nonlinear autoregressive neural network (NAR). ARIMA models are used to extract the linear correlations and the NAR neural network for modeling the residuals of ARIMA containing nonlinear components of the data.

Comparison: Single ARIMA model, ARIMA-NAR model and few other existing models which have been applied on the COVID-19 data in different countries are compared based on performance evaluation parameters.

Result:The hybrid combination displayed significant reduction in RMSE(16.23%), MAE(37.89%) and MAPE (39.53%) values when compared with single ARIMA model for daily observed cases. Similar results with reduced error percentages were found for daily reported deaths and cases of recovery as well. RMSE value of our hybrid model was lesser in comparison to other models used for forecasting COVID-19 in different countries.

Conclusion: Results suggested the effectiveness of the new hybrid model over a single ARIMA model in capturing the linear as well as nonlinear patterns of the COVID-19 data.

Keywords: Hybrid Model, Forecasting, COVID-19, ARIMA, NAR

1. INTRODUCTION

The novel coronavirus, COVID-19 (SARS-CoV-2), which was first reported in Wuhan, China, after the outbreak of exceptional pneumonia in late 2019, has already infected over 5.6 million people and caused more than three fifty thousand deaths worldwide [1]. Surpassing the fatalities caused by previous outbreaks such as severe acute respiratory syndrome coronavirus (SARS) [2,3], and middle east respiratory syndrome (MERS) [4,5], COVID-19 has been characterized by the world health organization (WHO) as a global pandemic [6]. The virus, which is assumed to be of zoonotic origin [7,8], has spread rapidly with a transmission rate of around 1.4 to 2.5 [9].

Therefore, to curb the outbreak, the nationwide lockdown has been observed in more than two hundred countries and in India. Table 1 shows the phases of lockdown conducted in India.

Table 1: Depiction of lockdown Phases

Lock Down Phases	Dates	Number of Cases	Days	Increase Percentage
Phase 0	22 January-24 March	2872	58	-
Phase 1	25th march-14april	10951	21	281.3%
Phase 2	15 april-3 may	31118	19	184.16%
Phase 3	3 may-17 may	53193	12	70.93%

COVID-19 first appeared in India in Kerala back in late January, where the patient had a recent travel record to Wuhan, China. Initially, the transmission was slow, and the virus could infect very few people within Kerala only. However, the number of cases started rising again in mid-march after the pandemic hit western Europe, and after that, strict lockdown measures were observed throughout the nation. Fig. 1 depicts the cumulative rise in no of cases in India under different phases of lockdown.

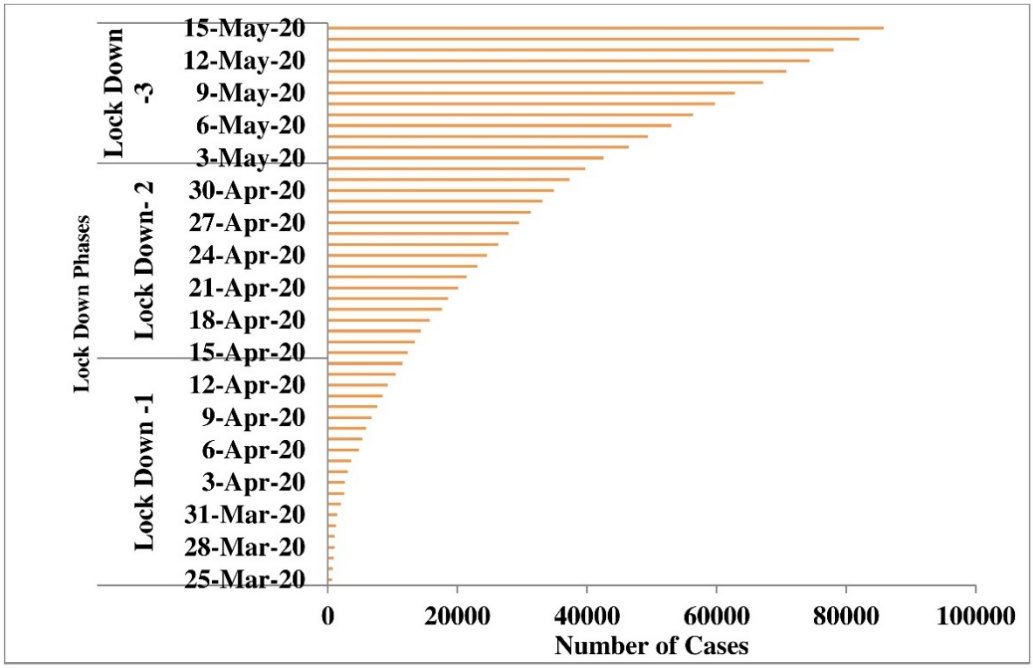

Fig.1: Confirmed cases of COVID-19 in India during Lockdown

India is the second-most populous country in the world after China. A slight negligence in constraining the pandemic can lead to unprecedented panic and widespread loss of trade, economy, outsourcing workforce, manufacturing, and other services all over the world. For all these, it is essential to have a proper strategy for combating the epidemic. In the current situation of unavailability of an adequate cure of the disease, having short term forecasts of the spread can provide state authorities with a realistic estimate of the magnitude of the outbreak for the coming weeks.

However, despite all the intervention strategies implemented by state authorities, the curve has jumped exponentially (fig. 2). Presently, the highest no of cases is observed in the United States; however, the curve is abruptly rising in Russia, India, and South American countries like Brazil.

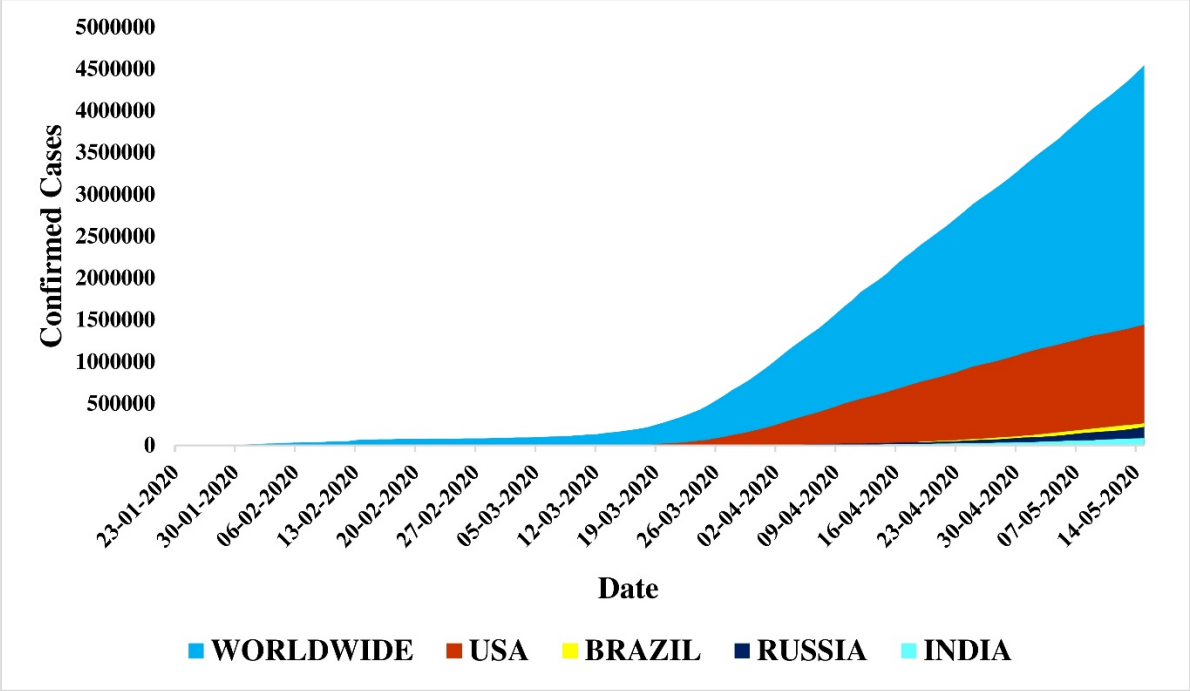

Fig.2: Total Confirmed cases of COVID-19 Worldwide from Jan 22 to May 15, 2020[1]

Time-series forecasting during epidemics has been regarded as an essential tool in the past for containing the spread of contagious diseases like ebola, influenza, etc.[10-16]. Timing plays a critical role in an epidemic, and from the very beginning, an exceptional level of monitoring is required to curb the spread. Several studies have shown that proper analysis of such outbreaks can contribute substantially in devising the right course of action in due time [17, 18]. In this connection, a standard model often used for analyzing the trend of an epidemic, 'susceptible–exposed–infectious–resistant' (SEIR), has been applied recently for analyzing COVID-19 cases in various countries [19-27].

Although such mathematical models are useful in epidemic analysis, they are based on coarse policies that are subject to bias [28]. Therefore researchers have subsequently proposed alternate forecasting models involving machine learning algorithms like LSTM, SVR, ARIMA, and few others for forecasting COVID-19 cases in different countries [29-43].

However, among all these forecasting models, ARIMA is most popular [44-46]. ARIMA works with an underlying assumption that the present data is linearly related to past observed values and errors. However, previous pandemics have often shown complex and nonlinear patterns with time, and therefore a linear approach might not yield the best results. Artificial Neural Networks (ANN) have emerged as one of the most successful methods to overcome this limitation of non-linearity [47-50]. However, ANN models are not capable of capturing both linear as well as nonlinear features of the time series equally well [51], and thus several hybrid methodologies have been developed [52-55]. Zhang [56] proposed a combination of ARIMA and NAR (Non-linear Auto-Regressive) Neural Network on some well-known datasets. Wang et al. [57] also implemented a similar model for forecasting tuberculosis cases in China. The same approach was opted by Benmouiza et al. in [58] for small-scale solar radiation forecasting. Most of the hybrid models were successful in improving the prediction accuracy as compared to the individual alternatives of those models. Therefore, the study of a hybrid model having capabilities of

modeling both linear and nonlinear time-series for COVID-19 could be capable of better forecasting.

With this motivation, we develop an ensemble model combining ARIMA and NAR models for predicting future cases of COVID-19 in India and then compare the results produced by the hybrid model with the regular one.

The organization of the rest of the paper is as follows: In Section 2, we discuss the methods for forecasting future COVID-19 cases along with the overall flow of the work. The implementation of these methods, along with a comparative analysis, is described in section 3. Section 4 holds a discussion, and Section 5 depicts the conclusion.

2. SYSTEM DESCRIPTION

Figure 3 depicts the workflow involved. In section 2.1, COVID-19 time-series data sources are mentioned. Section 2.2 describes our proposed ensemble model. A pictorial description of the same is presented in fig. 4. First we implement ARIMA model and analyze its results. Then to further improvise its results, a hybrid combination of ARIMA-NAR was developed. A comparison is made using performance evaluation parameters amongst these models. The section ends with a brief description of the accuracy estimation parameters in 2.3. All the ARIMA and NAR models are built in MATLAB v. 9.4.0.813654 (R2018a) using the Econometric Modeller Toolbox and Neural Net Time Series Toolbox respectively.

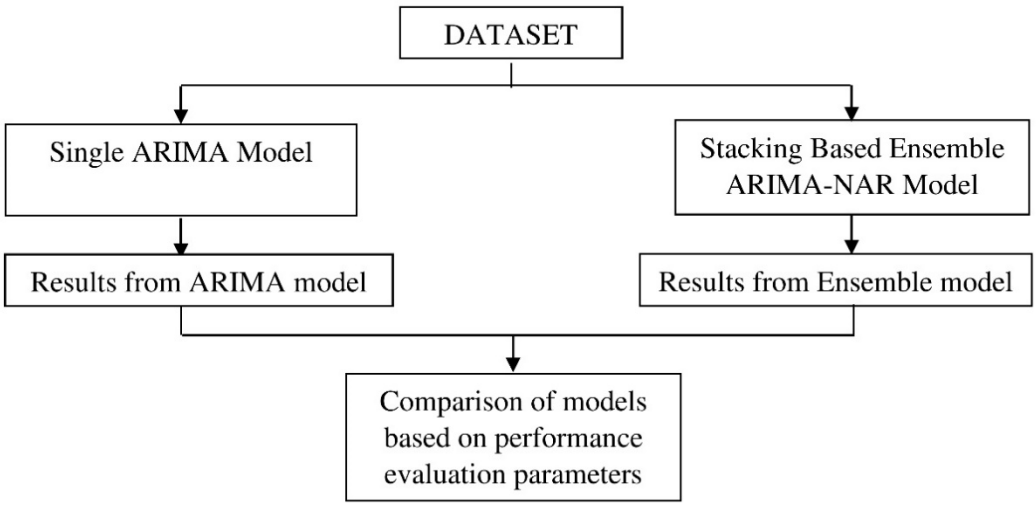

Fig.3: Flow chart indicating various steps involved for forecasting.

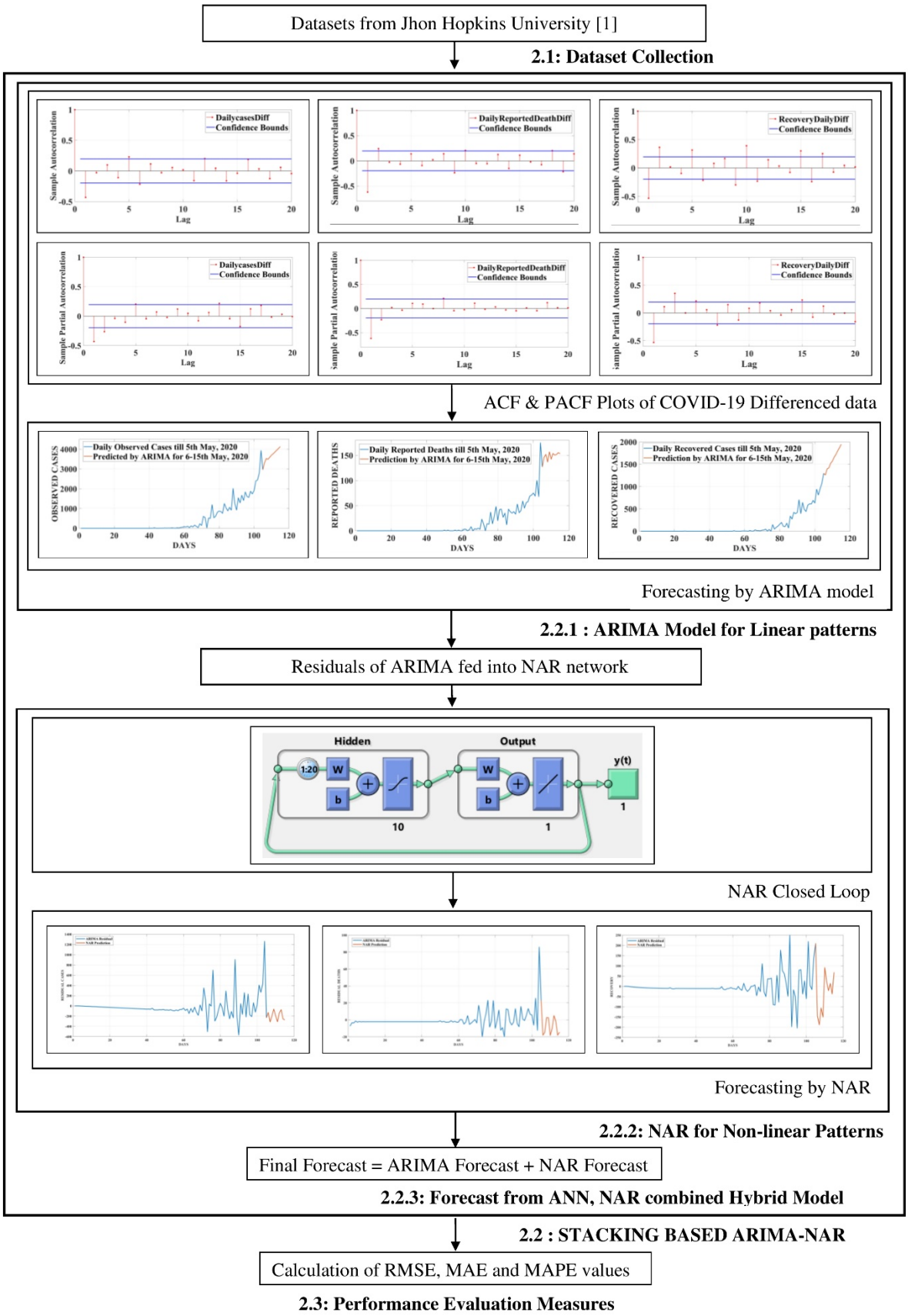

Fig.4: Pictorial description of the stack based ensemble ARIMA model

2.1. Data set collection

The cumulative count of confirmed cases, reported deaths and recovered cases of COVID-19 were taken from the official COVID-19 Data Repository of the Jhon Hopkins University [1] and for our study, we formulated the data in Microsoft Excel to obtain the respective cases on a daily basis betweenJanuary 22, 2020 and May 15, 2020 for India.

2.2. Stacking based ARIMA-NAR Model

Stacking based models basically use predictions from multiple models to build a new one. In this study, we utilize ARIMA models for extracting the linear relationships of the data and NAR neural network for the non linear patterns. Figure 4 gives a step wise explanation for the ARIMA-NAR ensemble model. First in 2.2.1, we describe the working of the ARIMA model. Next, section 2.2.2 talks about the NAR neural network and finally the contribution of both the models in making the final forecast is realized in section 2.2.3.

2.2.1. ARIMA Model for linear patterns

The econometric model, ARIMA was first presented by Box & Jenkins in 1970 [59]. The model is generally favored for its flexibility to various types of time-series data and its predicting accuracy.

ARIMA is a combination of A.R. and M.A. models, along with differencing. In Autoregressive models (A.R.), predictions are based on past values of the time-series data, and in Moving Average models (MA), prior residuals are considered for forecasting future values. The underlying process could be written as:

$$\boldsymbol{A_t = \theta_0 + \phi_1 A_{t-1} + \phi_2 A_{t-2} + \cdots + \phi_a A_{t-a} + E_t - \theta_1 E_{t-1} - \theta_2 E_{t-2} - \cdots - \theta_C E_{t-C}} \qquad (\mathbf{1})$$

Here, $\boldsymbol{A_t}$ is the actual observed value at time $\boldsymbol{t}$ and $\boldsymbol{E_t}$ is random error. $\boldsymbol{\phi_i}(i = 1,2, \dots, a)$ and $\boldsymbol{\theta_j}(j = 0,1,2, \dots, c)$ are model parameters where a and c denote order of the model. Random errors are generally independent and identically distributed with zero mean and constant variance.

In simpler terms, it represented as ARIMA (a, b, c) where 'a' denotes the order of A.R. model, 'b' is the differencing degree, 'c' is the order of the M.A. model. All these mentioned parameters of ARIMA model are determined in three iterative steps of model recognition, parameter selection and model verification.Since ARIMA models are generally suitable for stationary time series, so firstly in the identification step, stationarity of the time series is checked. If the series is not stationary, then differencing can be applied to make it stationary. After stationary tests, in the second step, appropriate parameters for the A.R. and M.A. models are selected for fitting based on Autocorrelation function (ACF) and Partial Autocorrelation Function (PACF) plots of the stationary data. In the final step, the goodness of the fit is verified by Akaike's Information Criterion (AIC) and Bayesian information criterion (BIC). These three steps are repeated untill a satisfactory model is achieved which is then used for forecasting.

2.2.2. NAR Neural Network for Nonlinear patterns

An artificial neural network (ANN) is an intuitive mapping structure represented by a mathematical model simulated around the biological nervous system. It is equipped with the ability to comprehend dynamic nonlinear time series patterns and arbitrary functions of all sorts. An ANN processes information by combining various neurons connected in a network of weighted links and then gives the output by computing certain activation functions that can be expressed in mathematical terms as mentioned:

$$\mathbf{Z} = \mathbf{f}\left(\mathbf{b} + \sum_{\mathbf{i}} \mathbf{w_i x_i}\right), \tag{2}$$

Where f is the activation function, b is the bias of neuron, w_i represents the weight, x_i input, and Z is the output.

Nonlinear autoregressive neural network (NAR) is a well-known ANN for modeling dynamic systems and predicting future values in a nonlinear time series [56-58]. It is based on the architecture of a recurrent neural network having embedded memory with feedback connections. The general equation of a NAR model could be defined as:

$$\boldsymbol{\widehat{Z}(t) = fx\left(Z(t-1) + Z(t-2) + \cdots + Z(t-n)\right)}, \tag{3}$$

Here, $\boldsymbol{fx}$ represents the nonlinear function, and the previous $\boldsymbol{n}$ output values determine the future values.

Among multiple architectures in a NAR model, the close loop network is widely used for multi-step ahead forecasting.

$$\boldsymbol{\widehat{Z}(t+s) = fx\left(Z(t-1) + y(t-2) + \cdots + y(t-n)\right)}, \tag{4}$$

Here, s denotes number of future points.

2.2.3. Forecast from ANN, NAR combined Hybrid Model

Although ARIMA and ANN both are potent methods for time-series forecasting, they have their own limitations. ARIMA models have achieved success in linear problems, whereas NAR models are more suitable for nonlinear domains [56-58]. While dealing with a real-world problem, it is challenging to ascertain all the characteristics of data, and therefore they study of a hybrid model having capabilities of modeling both linear and nonlinear time-series is essential.

In general, a time-series contains both linear autocorrelation structure as well as nonlinear components, and it could be written as:

$$\boldsymbol{Z_t = L_t + N_t}, \tag{5}$$

Where, Z_t is the original time-series data, L_t denotes the linear component, and N_t the nonlinear part at time t. The hybrid methodology is carried out in two steps. First, the linear component is modeled using ARIMA such that the residuals left after modeling will contain only the nonlinear relationship. If we can denote the residuals left by ARIMA at time t as R_t, then we get,

$$\boldsymbol{R_t = Z_t - \hat{L}_t}, \tag{6}$$

Where, $\hat{L}_t$ denotes forecasted valuesat time t by the ARIMA model.

Residual diagnosis plays a vital role in checking the sufficiency of ARIMA models. Although an ARIMA model is considered sufficient if the residuals left after fitting display no linear correlation structures, residual analysis cannot detect the presence of any significant nonlinear patterns in the data. Thus, by modeling the residuals using ANNs, nonlinear patterns can be realized. So, for the second step, the residuals are modeled to a NAR neural network with n input nodes as follows:

$$\boldsymbol{R_t = fx\,(R_{t-1}, R_{t-2}, \dots, R_{t-n}) + \epsilon_t,} \qquad (7)$$

Where,fx represents the nonlinear function evaluated by the NAR model and the leftover error is denoted by ϵ_t such that the final prediction can be equated as:

$$\boldsymbol{\hat{Z}_t = \hat{L}_t + \hat{N}_t,} \qquad (8)$$

Where, $\hat{Z}_t$ denotes the final predicted values at time t, and equation (7) is represented as $\hat{N}_t$, the forecast value of residuals.

The ARIMA-NAR combination thus exploits the strength of ARIMA as well as ANN models for capturing linear as well as nonlinear patterns.

Zhang [56] and Granger [60] have further pointed out the importance of the subjective selection of component models while building a hybrid model, as sometimes a combination of sub-optimal models can yield better forecasts for the hybrid model than that of the optimal ones.

2.3. Performance Evaluation Measures

In general, the performance of any forecasting model is determined by comparing the actual values with the predicted ones, and three standard methods for evaluation are:mean absolute percentage error (MAPE), root mean square error (RMSE) and mean absolute error (MAE). The optimum prediction model can thus beselected based on these performance measures.

$$\boldsymbol{RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(Z_t - \hat{Z}_t\right)^2},} \qquad (9)$$

$$\boldsymbol{MAE = \frac{1}{n}\sum_{t=1}^{n}\left|Z_t - \hat{Z}_t\right|,} \qquad (10)$$

$$\boldsymbol{MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{\left|Z_t - \hat{Z}_t\right|}{Z_t},} \qquad (11)$$

3. RESULTS

A total of 85,784 cases of novel coronavirus were reported throughout India along with 2,753 deaths and 30,258 cases of recovery till May 15, 2020. Figure 5 shows the number of cases observed on a daily basis, daily reported deaths and daily recovered cases in India between January 22 and May 15, 2020. We utilize the data from Jan 22 to May 5, 2020 for training purpose and then test the respective models for 6-15 May, 2020 for all three datasets.

Section 3.1 describes the major steps involved in selecting the ARIMA model along with the results achieved at each step for all the three datasets. In section 3.2, we show the training steps of NAR neural network along with several plots verifying the model adequacy. Finaly, section 3.3 ends with forecasting results of both the models along with a comparitive analysis in terms of RMSE, MAE and MAPE values.

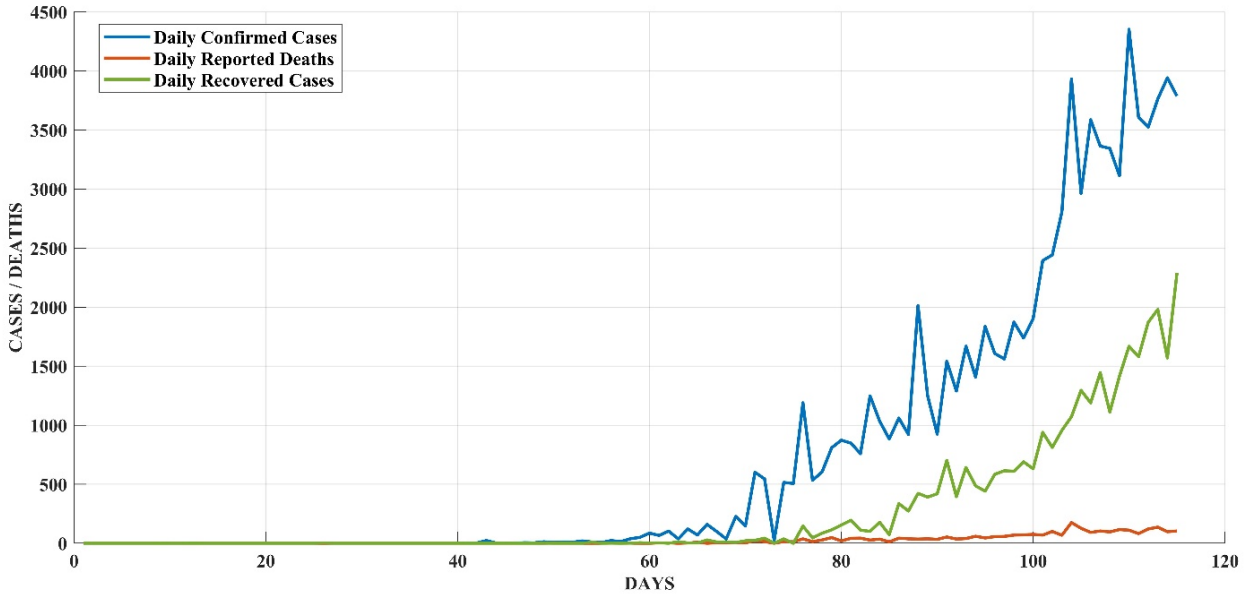

Fig.5: Daily observed cases, reported deaths and recovered cases of COVID-19 in India till May 15, 2020

3.1. Selecting the best-fit ARIMA model

We started with the first step of model identification. Augmented Dickey-Fuller (ADF) unit root test [61] is a widely used method for checking stationarity of time-series data. We performed the ADF unit root test with a significance level of 0.05 on daily observed cases, daily reported deaths, and daily cases of recovery in India. With ADF test results, it was confirmed that all the three time-series data were not stationary and needed differencing. After differencing of all the datasets, ADF tests were repeated again to show that the data had became stationary (table. 2-a, 2-b, 2-c). This differencing also suggested a probable value of 'b=1' for the ARIMA model (a, b, c). After stationary tests, appropriate parameters for the A.R. and M.A. models are choosen based on ACF and PACF plots of the differenced data. We analysed the ACF and PACF plots of the differenced data for all the three datasets (fig. 6-11), and proposed a series of candidate models for fitting (table. 3-a, 3-b, 3-c). Finaly, after fitting these models, the goodness of the fit was verified by AIC and BIC values and thus the optimum model was selected for forecasting.

Table 2. (a) Augmented Dickey-Fuller Test, Null Hypothesis: India Daily Confirmed Cases contains a unit root

Time Series	*Null Rejected*	*P-Value*	*Test Statistic*	*Critical Value*
IndiaDaily	False	0.999	3.0774	-1.9443
IndiaDaily1st difference	True	0.001	-10.1548	-1.9443

Table 2. (b) Augmented Dickey-Fuller Test, Null Hypothesis: India Daily Reported Deaths contains a unit root

Time Series	*Null Rejected*	*P-Value*	*Test Statistic*	*Critical Value*
IndiaReported Deaths	False	0.999	4.8635	-1.9443
IndiaReported 1st difference	True	0.001	-7.3662	-1.9443

Table 2. (c) Augmented Dickey-Fuller Test, Null Hypothesis: India Daily Recovery Cases contains a unit root

Time Series	*Null Rejected*	*P-Value*	*Test Statistic*	*Critical Value*
India Recovered	False	0.999	6.5746	-1.9443
India Recovered 1stDifference	True	0.001	-6.6969	-1.9443

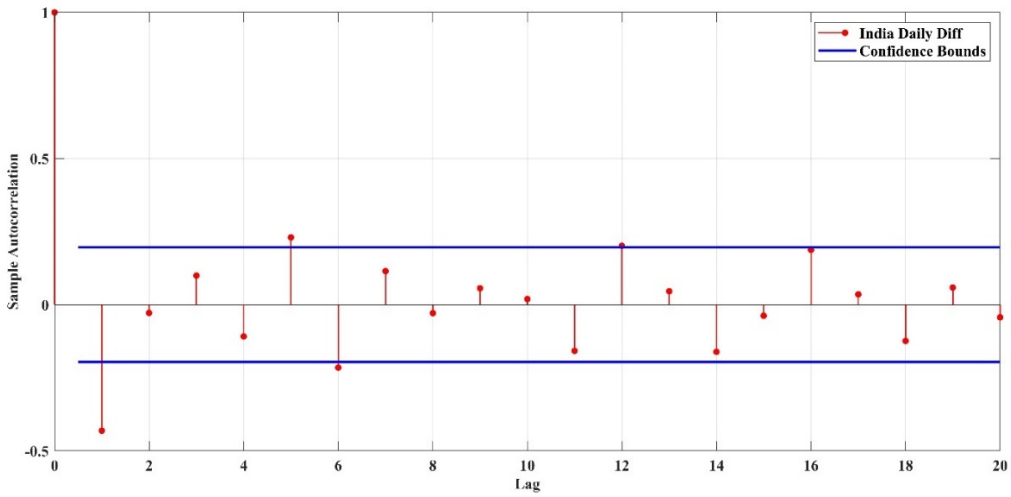

Fig.6: Auto-correlation Function of India_Daily_1stDifference

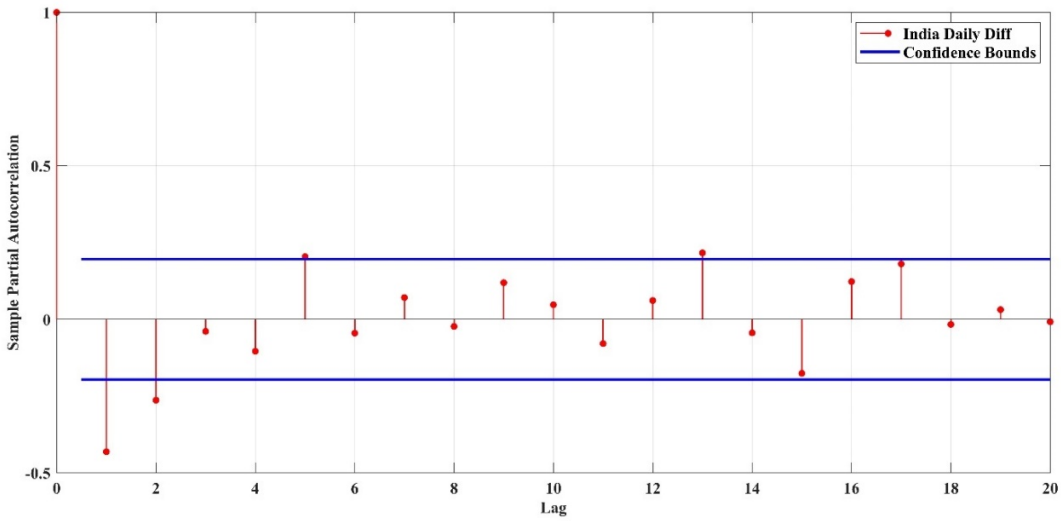

Fig.7: Partial Auto-correlation Function of India_Daily_1st Difference

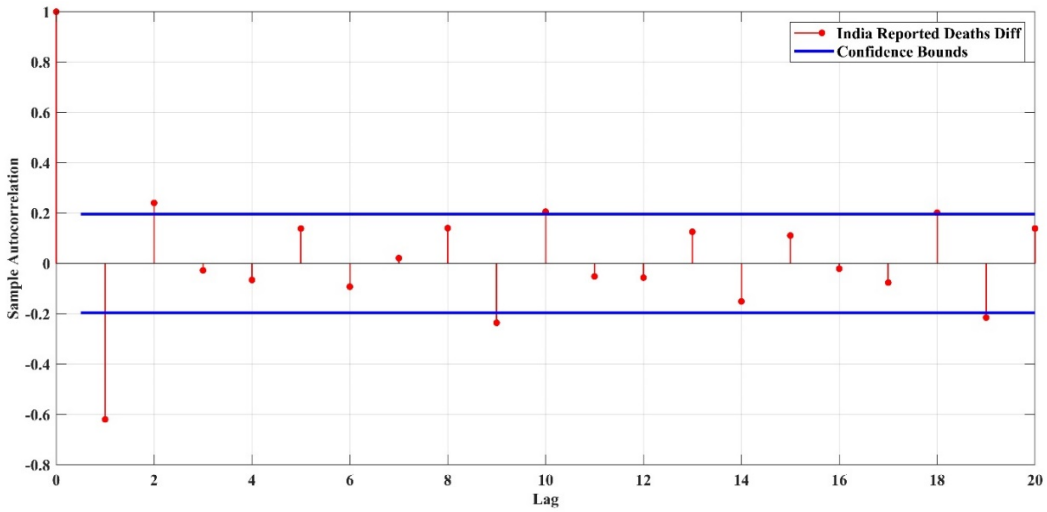

Fig.8: Auto-correlation Function of India_Deaths_1st Difference

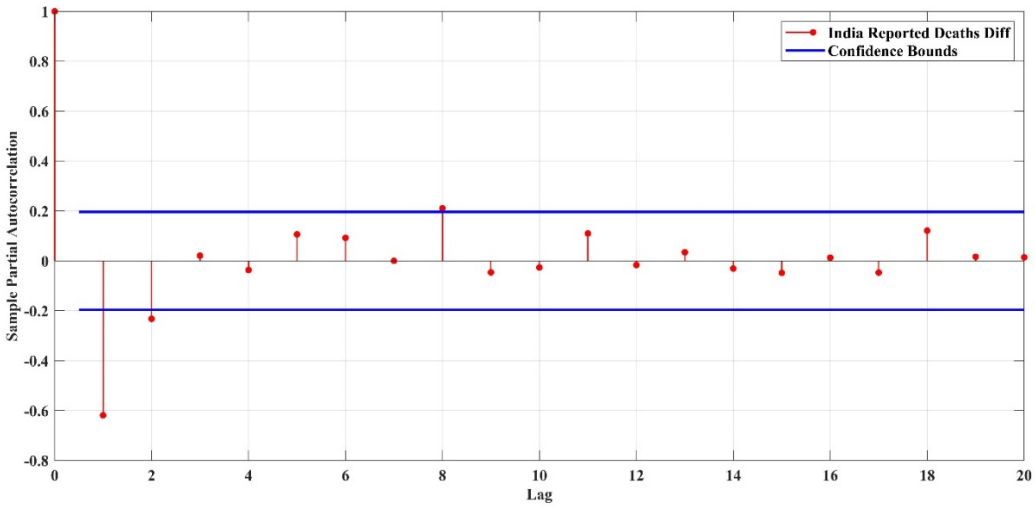

Fig.9: Partial Auto-correlation Function of India_Deaths_1st Difference

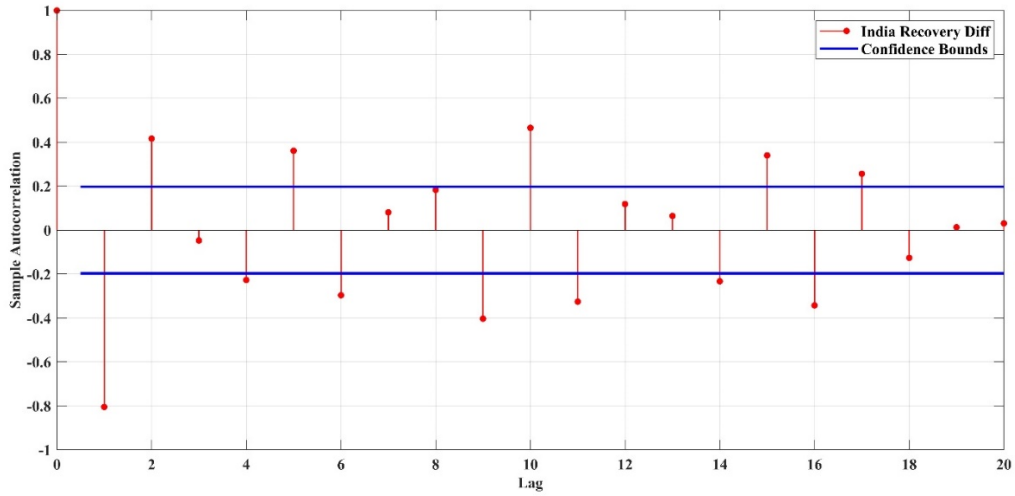

Fig.10: Auto-correlation Function of India_Recovery_1st Difference

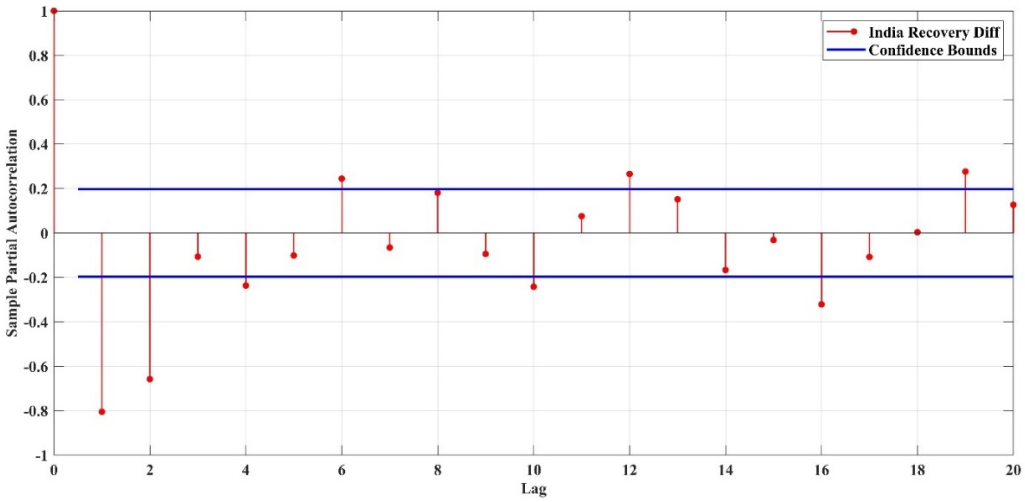

Fig.11: Partial Auto-correlation Function of India_Recovery_1st Difference

Table 3. (a) Goodness of Fit Statistics for India Daily Confirmed Cases Time Series generated by ARIMA

ARIMA Model	*ARIMA (2,1,0)*	*ARIMA (2,1,1)*	*ARIMA (2,1,2)*	*ARIMA (3,1,1)*
AIC	1448.0586	**1441.5167**	1449.0606	1443.2667
BIC	1461.1835	**1454.5923**	1464.8104	1458.9574

Table 3. (b) Goodness of Fit Statistics for India Daily Reported Deaths Time Series generated by ARIMA

ARIMA Model	*ARIMA (2,1,1)*	*ARIMA (2,1,2)*	*ARIMA (3,1,1)*	*ARIMA (3,1,2)*

AIC	814.706	**794.3605**	813.4623	802.5842
BIC	827.8309	**810.0512**	829.2121	815.6598

Table 3. (c) Goodness of Fit Statistics for India Daily Recovered Cases Time Series generated by ARIMA

ARIMA Model	***ARIMA (2,1,1)***	***ARIMA (2,1,2)***	***ARIMA (3,1,1)***	***ARIMA (3,1,2)***
AIC	1166.6708	**1158.8243**	1162.7776	1160.8243
BIC	1179.7957	**1171.998**	1181.0137	1176.5741

3.2. Constructing the Hybrid model

The residual data of the ARIMA forecast from all the three-time series data for January 22 to May 5, 2020, were fed to the NAR network as respective inputs. After dividing the data into training, validation, and testing, lag value and count of hidden layer neurons are determined using hit and trial methods. Several weight optimising algorithms are used in an ANN for adjusting the weight values, and the 'Neural Net Time Series Toolbox' in MATLAB provides three sets of such algorithms, namely Levemberg–Marquardt [62], Bayesian Regularization [63] and scaled conjugate gradient [64]. Low MSE and higher R values for multiple iterations of the respective subsets accounted for selection the optimum NAR model. The error autocorrelation plot for all the datasets (figure 12-14) is also used for verifying the adequacy of the model. Beale MH et al. [65] in their study have shown that for an ideal prediction model, except for the non-zero value at zero lag, most of the error values should come inside the confidence limits which is true in our case. After the training is finished, all the synaptic weights are saved, and the model is ready for prediction.

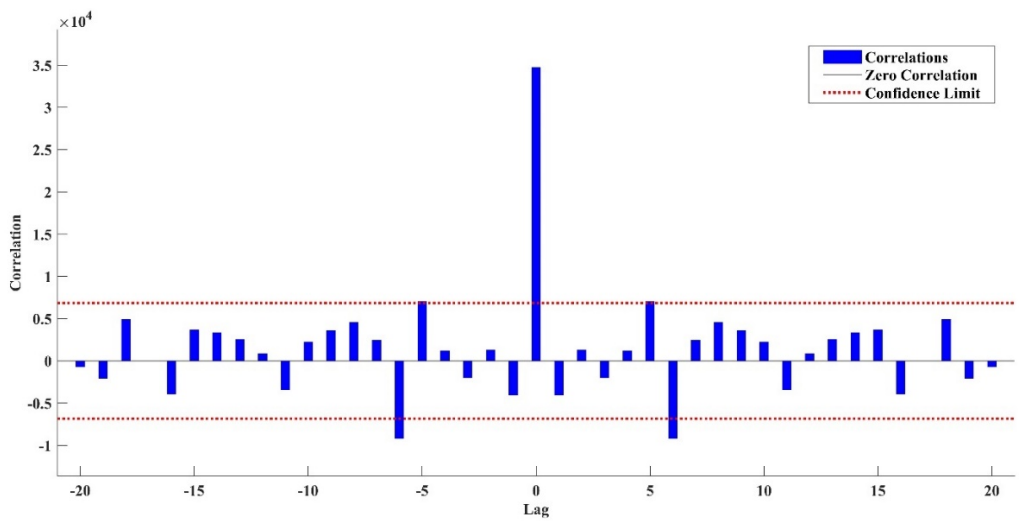

Fig.12: Error auto-correlation plot for NAR-Residual of Daily Confirmed Cases

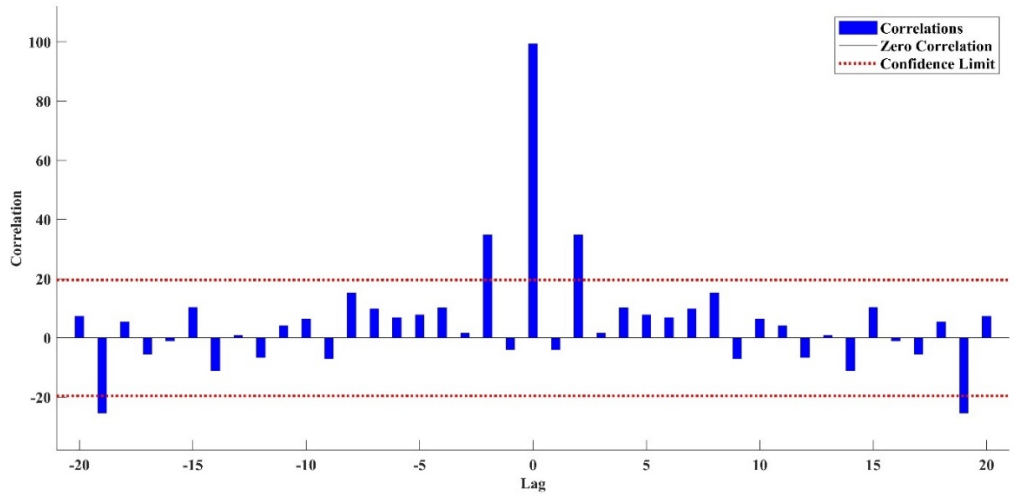

Fig.13: Error auto-correlation plot for NAR-Residual of Daily Reported Deaths

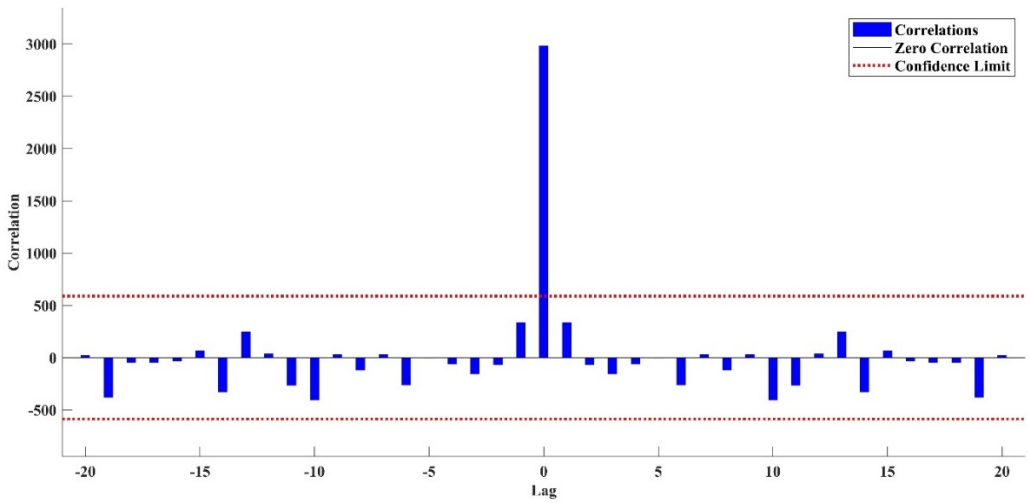

Fig.14: Error auto-correlation plot for NAR-Residual of Daily Recovered Cases

3.3. Forecasting and Comparison of Models

After generating respective fitting curves and residual plots, the ARIMA model is used for prediction on all the three-time series data between May 6-15, 2020. Figure (15-20) respectively show the prediction of future cases and residuals by the ARIMA and NAR neural network for daily observed cases, reported deaths, and daily recovered cases. The final forecasting is done by combining the separate prediction values of ARIMA and NAR models. RMSE, MAE and MAPE values are calculated for the predictions made by single ARIMA model and the ARIMA-NAR combined model for all the three datasets (table 4-a, 4-b, 4-c).

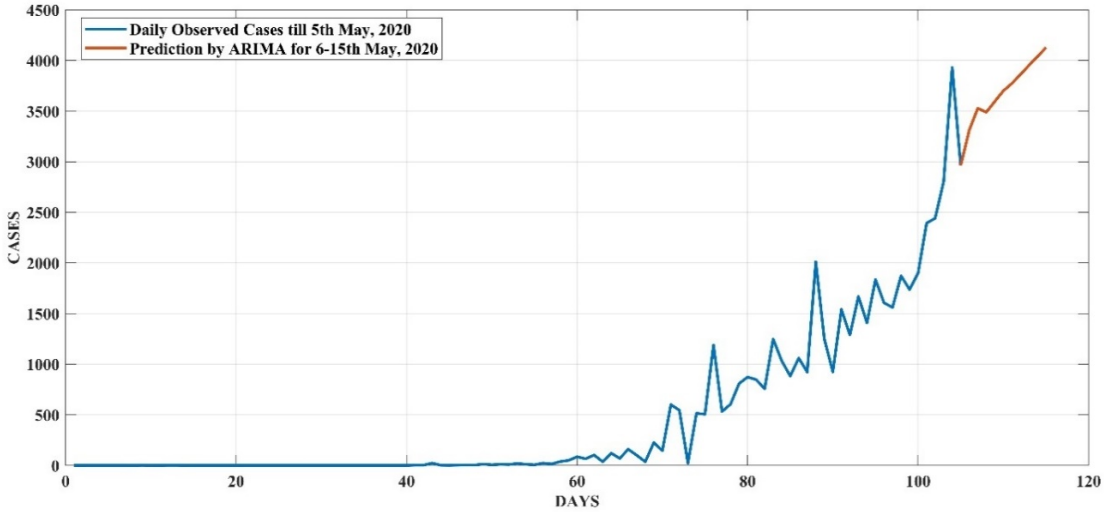

Fig.15: Prediction by ARIMAmodel for Daily observed casesin India between 6th and 15th May, 2020

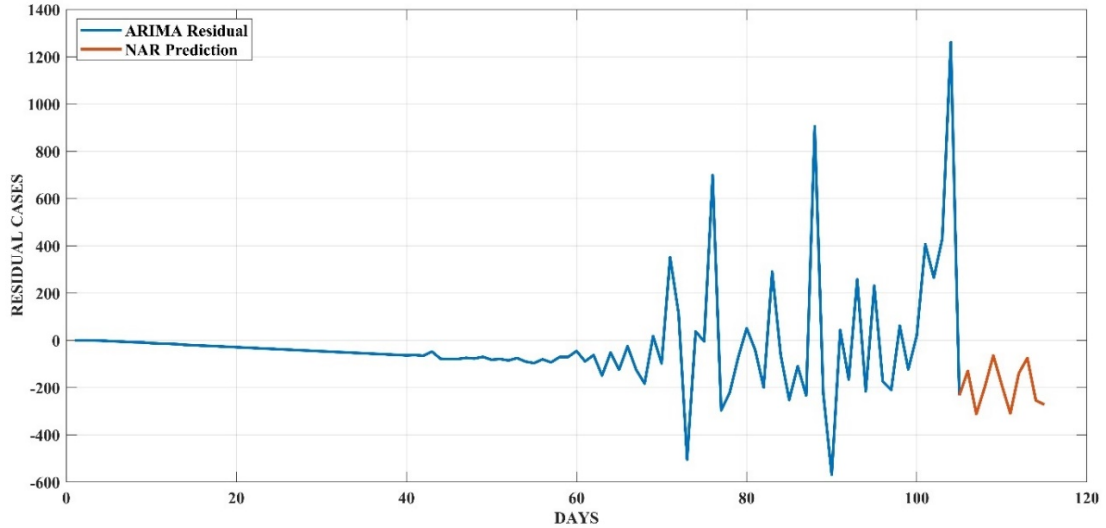

Fig.16: Prediction of residual error by NAR model for daily observed cases in India between 6th and 15th May, 2020

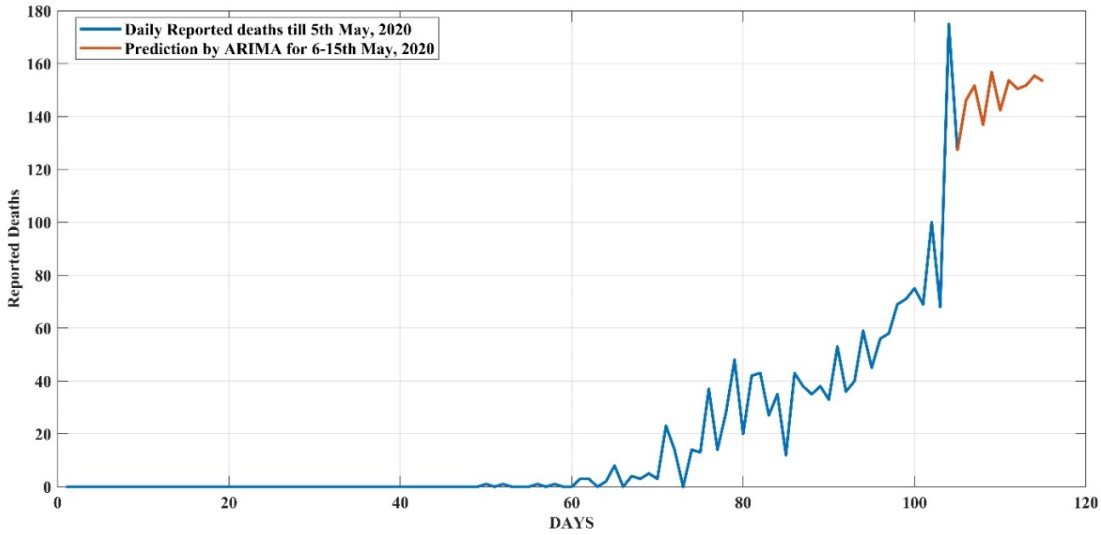

Fig.17: Prediction by ARIMAmodel for Daily Reported deathsin India between 6th and 15th May, 2020

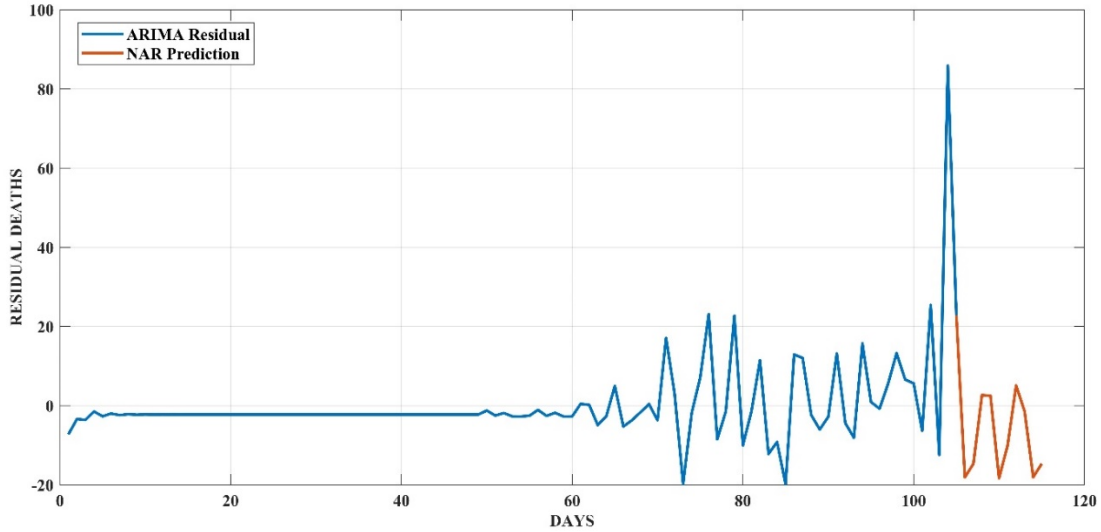

Fig.18: Prediction of residual error by NAR model for daily reported deaths in India between 6th and 15th May, 2020

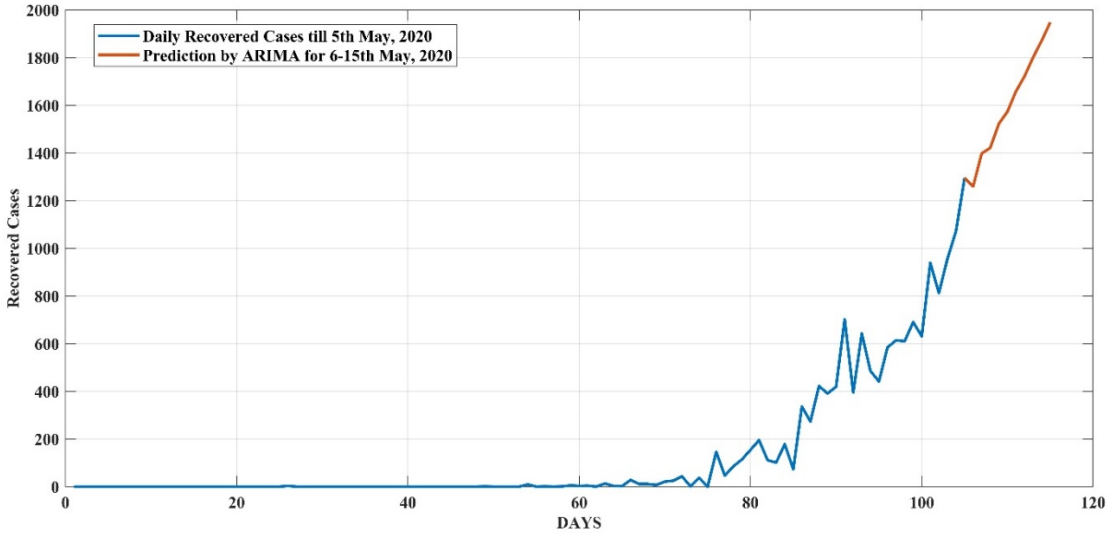

Fig.19: Prediction by ARIMAmodel for Daily recovered cases in India between 6th and 15th May, 2020

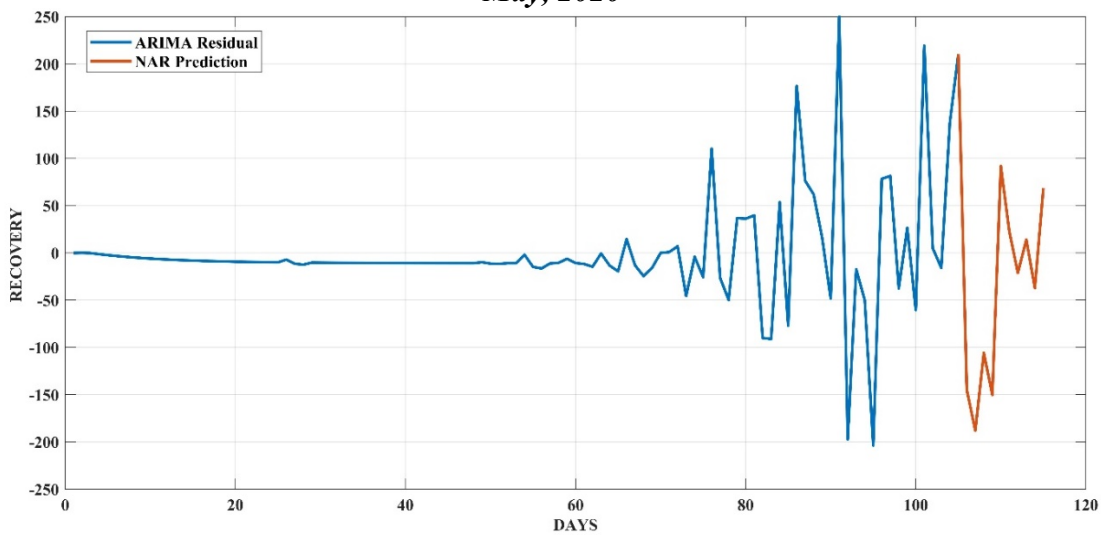

Fig.20: Prediction of residual error by NAR model for daily recovered cases in India between 6th and 15th May, 2020

Table 4. (a) Prediction Accuracy evaluation for daily observed cases in India between 6th and 15th May, 2020

Model	*RMSE*	*MAE*	*MAPE*
Single Arima	329.4373	284.9	0.078045
Hybrid Arima	**275.9648**	**176.9298**	**0.047187**

Table 4. (b) Prediction Accuracy evaluation for daily reported deaths in India between 6th and 15th May, 2020

Model	*RMSE*	*MAE*	*MAPE*

Single Arima	46.3923	43.8708	0.440245
Hybrid Arima	**37.79482**	**35.3597**	**0.3532**

Table 4. (c) Prediction Accuracy evaluation for daily recovered cases in India between 6th and 15th May, 2020

Model	***RMSE***	***MAE***	***MAPE***
Single Arima	198.0642	168.1494	0.106625
Hybrid Arima	**177.6032**	**153.3469**	**0.09673**

As seen in table 4-a, 4-b and 4-c, hybrid ARIMA's performance provide more adequate results. The RMSE, MAE and MAPE value of the hybrid combination for daily observed cases are 275.9648 (16.23% reduction), 176.9298 (37.89% reduction), 0.047187 (39.53% reduction). Regarding daily reported deaths, cases of recovery and cumulative confirmed cases similar results were found with reduced error percentages.

4. DISCUSSION

The current COVID-19 outbreak has brought forward a major challenge for healthcare sector all over the world. After witnessing a catastrophic rise in the number of COVID-19 cases in USA and western Europe, a proper strategy for epidemic control in a densely populated country like India has become priority and to implement control measures in due time, forecasting of future cases is certainly essential. Several forecasting models have been proposed in recent months for predicting future cases of COVID-19 in different countries. Most of the forecasting work has been done using standard ARIMA models which are popular for their statistical properties in building models.

Generally, a time series compromises of linear as well as nonlinear patterns and the existing trend of COVID-19 over last few months clearly depicted nonlinear patterns (fig 5). While ARIMA models have proven quite useful for linear time-series, they cannot extract nonlinear patterns sufficiently. On the other hand, NAR, a powerful class of ANN has displayed favourable characteristics for modelling nonlinear time-series. However, ANN models have their own limitations in equally capturing both the linear and nonlinear patterns. Therefore, a hybrid approach that utilizes ARIMA and ANN models together is proposed in the present study.

Our study highlighted the key point of analysing linear and nonlinear patterns using separate models in context of a time series forecasting. Three separate datasets (fig 5) of daily confirmed cases of COVID-19 in India, reported deaths and cases of recovery were respectively trained on both the models for a duration of over hundred days between January 22 to May 5, 2020. All the datasets were first tested for stationarity by ADF unit root test (table 2-a, 2-b, 2-c) before fitting to ARIMA models. A series of candidate models were proposed afterwards based on the ACF and PACF plots (fig 6-11) of the differenced data. Next, using AIC and BIC values of the candidate models (table. 3-a, 3-b, 3-c), the best model was selected for training the respective datasets and subsequently the fitting curve and residual plot of all the three datasets were generated.

Further, for extracting the nonlinear patterns, the residuals left from the ARIMA models were fitted to the NAR neural network. Error Auto-correlation plots of the residuals from all the datasets (fig. 12-14) were used to verify the sufficiency of the NAR neural network. Both the models, ARIMA and NAR were then used to predict the future cases and residuals respectively for a duration of ten days between May 6 to May 15, 2020 (fig 15-20). The combination of prediction results from both these models were used as the final results for the hybrid model.

Our hybrid ARIMA model was able to capture the nonlinear patterns quite well which were left as residuals by the ARIMA model. On the basis of RMSE, MAE, and MAPE measures (eq. 9-11), we evaluated the prediction accuracy of both the models for all the three datasets. Reduced error as seen in table 4-a, 4-b and 4-c clearly advocate for the superiority of the proposed hybrid ARIMA model over a single ARIMA model. We further draw a rough comparison of our hybrid model with some of the existing ones that have been applied over covid-19 in table 5 where our model clearly out performs other existing models. This is in line with previous studies where a combination of ARIMA and NAR model has been explored as a possibility for producing better time-series forecasting results. Hence, the present study can be regarded as an authentic approach for time-series forecasting during pandemics.

Table 5. Comparison of our Hybrid model with few existing models over COVID-19 data in different countries

Author	***Dataset Duration***	***Country***	***Results***		
			Methods	***RMSE (Daily Confirmed Cases)***	***RMSE (Total Confirmed Cases)***
Al-Qaness et.al [41]	30 days	China	*ANN*	8750	NA
			KNN	12,100	
			SVR	7822	
			ANFIS	7375	
			PSO	6842	
			GA	7194	
			ABC	8327	
			FPA	6059	
			FPASSA	5779	
Ceylan et.al [42]	45 days	France	*ARIMA*	NA	971.9250
		Italy			1654.6600
		Spain			2031.1200
Punn et.al [32]	71 days	Worldwide	*SVR*	NA	27456.47
			DNN		163335.65
			LSTM		15647.64
			PR		455.92
Moftakhar et al. [43]	71 days	Iran	*ANN*	746.60	NA
			ARIMA	1539.43	

Proposed	**115 days**	**India**	***Hybrid Arima***	**275.9648**	**437.192193**

Acknowledgments

This work is under the project "DEVELOPMENT OF ENSEMBLE MODEL FOR PREDICTING TRENDS OF COVID-19". We thank Jhon Hopkins University [1] for publicly providing respective time-series data of confirmed cases, deaths and recovery for our research work.

5. CONCLUSION

In this paper, we presented a new hybrid model for COVID-19 time-series forecasting by combining an Auto-Regressive Integrated Moving Average (ARIMA) model with a Nonlinear Auto-Regressive (NAR) neural network. ARIMA models were used to capture the linear relationship from the time-series, and the residuals of the ARIMA model containing the nonlinear components were fitted by the NAR Model. The prediction accuracy of both the models were measured on the basis of Root Mean Squared Error, Mean Absolute Error, and Mean Absolute Percentage Error. With low values of RMSE, MAE, and MAPE, the combination of ARIMA-NAR models produced better prediction results as compared to the single ARIMA model. Therefore, the new hybrid model can be considered as a reliable tool for policymakers in predicting future cases of COVID-19 and devising proper strategies in due time.

Since forecasting requires a large amount of data for better prediction, a lack of sufficient data could be considered as the only limitation in this study.

List of Acronyms

ABC – Artificial Bee Colony

ACF – Auto-Correlation Function

ADF – Augmented Dickey-Fuller

AIC – Akaike's Information Criterion

ANFIS – Adaptive Neuro-Fuzzy Inference System

ANN – Artificial Neural Networks

AR – Auto-Regressive

ARIMA – Autoregressive Integrated Moving Average

BIC – Bayesian Information Criterion

COVID-19 – Coronavirus Disease -2019

DNN – Deep Neural Network

FPA – Flower Pollination Algorithm

FPASSA – Flower pollination algorithm Salp Swarm Algorithm

GA – Genetic Algorithm

GROOMS – Group of Optimized and Multisource Selection

IoT – Internet of Things

KNN – K-Nearest Neighbors

MA – Moving Average

MAE – Mean Absolute Error

MAPE – Mean Absolute Percentage Error

MERS – Middle East Respiratory Syndrome

ML – Machine Learning

NAR – Nonlinear Autoregressive

PACF – Partial Auto-Correlation Function

PR – Polynomial Regression

PSO – Particle Swarm Optimization

RMSE – Root Mean Square Error

SARS – Severe Acute Respiratory Syndrome

SARS-CoV-2 –Severe Acute Respiratory Syndrome Coronavirus 2

SEIR –Susceptible–Exposed–Infectious–Resistant

SES – Single Exponential Smoothing

SIRD – Suspected-Infected-Recovered-Dead)

SVR – Support Vector Regression

WHO – World Health Organization

WMA – Weighted Moving Average

Compliance with Ethical Standards:

Funding: No Funding Received.

Conflict of Interest: The authors declare that they have no conflict of interest.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

REFERENCES

[1] Johns Hopkins University Centre for Systems Science and Engineering, 2019. https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data.

[2] World Health Organization. 2004. Available at: https://www.who.int/ith/diseases/sars/en/ (accessed January 2020)

[3] Centres for Disease Control and Prevention. 2017. Available at: https://www.cdc.gov/sars/about/fs-sars.html (accessed January 2020)

[4] World Health Organization. 2019. Available at: https://www.who.int/emergencies/mers-cov/en/ (accessed January 2020)

[5] Oboho, Ikwo K., et al. "2014 MERS-CoV outbreak in Jeddah—a link to health care facilities." New England Journal of Medicine 372.9 (2015): 846-854.

[6] World Health Organization, 2020. Coronavirus disease 2019 (COVID-19): situation report, 51.

[7] Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020. [Epub ahead of print].

[8] Li Q, Guan X, Wu P, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. N Engl J Med 2020. [Epub ahead of print].

[9] Mahase, Elisabeth. "China coronavirus: what do we know so far?" (2020).

[10] W. Jia, X. Li, K. Tan, and G. Xie, "Predicting the outbreak of the hand-foot- mouth diseases in china using recurrent neural network," in 2019 IEEE International Conference on Healthcare Informatics (ICHI). IEEE, 2019, pp. 1–4.

[11] Shashvat, Kumar, RikmantraBasu, and Amol P. Bhondekar. "Application of time series methods for dengue cases in North India (Chandigarh)." Journal of Public Health (2019): 1-9.

[12] A. Forna, P. Nouvellet, I. Dorigatti, and C. Donnelly, "Case fatality ratio estimates for the 2013–2016 west African Ebola epidemic: application of boosted regression trees for imputation," International Journal of Infectious Diseases, vol. 79, p. 128, 2019.

[13] Shashvat, Kumar, et al. "Comparison of time series models predicting trends in typhoid cases in northern India." Southeast Asian Journal of Tropical Medicine and Public Health 50.2 (2019): 347-356.

[14] S.-L. Jhuo, M.-T. Hsieh, T.-C. Weng, M.-J. Chen, C.-M. Yang, and C. H. Yeh, "Trend prediction of influenza and the associated pneumonia in Taiwan using machine learning," in 2019 International Symposium on Intelligent Signal Processing.

[15] G. Machado, C. Vilalta, M. Recamonde-Mendoza, C. Corzo, M. Torremorell, A. Perez, and K. VanderWaal, "Identifying outbreaks of porcine epidemic diarrhoea virus through animal movements and spatial neighbourhoods," Scientific reports, vol. 9, no. 1, pp. 1–12, 2019.

[16] Kalipe, G., Gautham, V., Behera, R. K. (2018, December). Predicting malarial outbreak using Machine Learning and Deep Learning approach: A review and analysis. In 2018 International Conference on Information Technology (ICIT) (pp. 33-38). IEEE.

[17] Singh, R., Singh, R., Bhatia, A. (2018). Sentiment analysis using Machine Learning technique to predict outbreaks and epidemics. Int. J. Adv. Sci. Res, 3(2), 19-24.

[18] S. A. Abdulkareem, E.-W. Augustijn, T. Filatova, K. Musial, and Y. T. Mustafa, "Risk perception and behavioural change during epidemics: Comparing models of individual and collective learning,"PloS one, 2020.

[19] Kuniya T. Prediction of the Epidemic Peak of Coronavirus Disease in Japan, 2020. J Clin Med. 2020; 9 (3): E789. Published 2020 March 13. doi: 10.3390 / jcm9030789.

[20] Gupta, Rajan, et al. "SEIR and Regression Model based COVID-19 outbreak predictions in India." medRxiv (2020).

[21] Yuan, George Xianzhi, et al. "The Framework for the Prediction of the Critical Turning Period for Outbreak of COVID-19 Spread in China based on the iSEIR Model." Available at SSRN 3568776 (2020).

[22] C. Anastassopoulou, L. Russo, A. Tsakris, and C. Siettos, "Data-Based Analysis, Modelling and Forecasting of the novel Coronavirus (2019-nCoV) outbreak,"medRxiv, no. February, p. 2020.02.11.20022186, 2020, doi: 10.1101/2020.02.11.20022186.

[23] Joseph T. Wu, Kathy Leung, Mary Bushman, Nishant Kishore, Rene Niehus, Pablo M. de Salazar, Benjamin J. Cowling, Marc Lipsitch& Gabriel M. Leung. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. Nature Medicine (2020), March 19 2020.https://doi.org/10.1038/s41591-020-0822-7.

[24] J. T. Wu, Leung, K. & Leung, G. M. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet 2020; 395: 689–97. https://doi.org/10.1016/ S0140-6736(20)30260-9.

[25] Kiesha Prem, Yang Liu, Timothy W Russell, Adam J Kucharski, Rosalind M Eggo, Nicholas Davies. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. Lancet Public Health 2020. Published Online March 25, 2020.https://www.thelancet.com/journals/lanpub/article/PIIS2468-2667(20)30072-4/fulltext.

[26] X. Liu, Geoffrey Hewings, Shouyang Wang, Minghui Qin, Xin Xiang, Shan Zheng, Xuefeng Li. Modelling the situation of COVID-19 and effects of different containment strategies in China with dynamic differential equations and parameters estimation. medRxiv preprint doi: https://doi.org/10.1101/2020.03.09.20033498, 2020.

[27] Qianying Lin, Shi Zhao, Daozhou Gao, Yijun Lou, Shu Yang, Salihu S. Musa, Maggie H. Wang, Yongli Cai, Weiming Wang, Lin Yang, Daihai He. A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action. International Journal of Infectious Diseases (93) (2020), 211-216.

[28] J.L. Murray. Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator days and deaths by U.S. state in the next 4 months. MedRxiv. March 26 2020. doi:10.1101/2020.03.27.20043752.

[29] Elmousalami, H. H., &Hassanien, A. E. (2020). Day Level Forecasting for Coronavirus Disease (COVID-19) Spread: Analysis, Modelling and Recommendations. ArXiv preprint arXiv:2003.07778.

[30] Pal, Ratnabali, et al. "Neural network-based country wise risk prediction of COVID-19." arXiv preprint arXiv:2004.00959 (2020).

[31] Bandyopadhyay, Samir Kumar, and Shawni Dutta. "Machine Learning Approach for Confirmation of COVID-19 Cases: Positive, Negative, Death and Release." medRxiv (2020).

[32] Punn, Narinder Singh, Sanjay Kumar Sonbhadra, and Sonali Agarwal. "COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms." medRxiv (2020).

[33] Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M. Application of the ARIMA model on the COVID-2019 epidemic dataset. Data Brief. 2020; 29: 105340. Published 2020 Feb 26. doi: 10.1016 / j .dib.2020.105340.

[34] Ding, Guorong, et al. "Brief Analysis of the ARIMA model on the COVID-19 in Italy." medRxiv (2020).

[35] Perone, Gaetano. An ARIMA model to forecast the spread and the final size of COVID-2019 epidemic in Italy. No. 20/07. HEDG, c/o Department of Economics, University of York, 2020.

[36] Dehesh, T., Mardani-Fard, H. A., &Dehesh, P. (2020). Forecasting of COVID-19 Confirmed Cases in Different Countries with ARIMA Models. MedRxiv.

[37] Gupta, Rajan, and Saibal Kumar Pal. "Trend Analysis and Forecasting of COVID-19 outbreak in India." medRxiv (2020).

[38] Tandon, Hiteshi, et al. "Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future." arXiv preprint arXiv:2004.07859 (2020).

[39] Kumar, Pavan, et al. "Forecasting the dynamics of COVID-19 Pandemic in Top 15 countries in April 2020: ARIMA Model with Machine Learning Approach." medRxiv (2020).

[40] Shi, Z., & Fang, Y. (2020). Temporal relationship between outbound traffic from Wuhan and the 2019 coronavirus disease (COVID-19) incidence in China. MedRxiv.

[41] Al-Qaness, Mohammed AA, et al. "Optimization method for forecasting confirmed cases of COVID-19 in China." Journal of Clinical Medicine 9.3 (2020): 674.

[42] Ceylan, Zeynep. "Estimation of COVID-19 prevalence in Italy, Spain, and France." Science of The Total Environment (2020): 138817.

[43] Moftakhar, Leila, S. E. I. F. Mozhgan, and Marziyeh Sadat Safe. "Exponentially Increasing Trend of Infected Patients with COVID-19 in Iran: A Comparison of Neural Network and ARIMA Forecasting Models." Iranian Journal of Public Health 49 (2020): 92-100.

[44] Zhang, Y.; Yang, H.; Cui, H.; Chen, Q. Comparison of the Ability of ARIMA, WNN and SVM Models for Drought Forecasting in the Sanjiang Plain, China. Nat. Resour. Res. 2019, 29, 1447.

[45] Zhang X, et al. Applications and comparisons of four time series models in epidemiological surveillance data. PLoS ONE 2014; 9: e88075.

[46] Li Q, et al. Application of an autoregressive integrated moving average model for predicting the incidence of haemorrhagic fever with renal syndrome. American Journal of Tropical Medicine and Hygiene 2012; 87: 364–370.

[47] Adly A.A., et al. Utilizing neural networks in magnetic media modelling and field computation: a review. Journal of Advanced Research 2014; 5: 615–627.

[48] Haykin S. Neural networks: a comprehensive foundation. 2nd ed. Prentice Hall; 1998.

[49] Ljung L. System identification: theory for the user. 2nd ed. Prentice Hall PTR; 1998.

[50] Connor JT, et al. Recurrent neural networks and robust time series prediction. IEEE Transactions on Neural Networks 1994; 5: 240–254.

[51] Taskaya-Temizel T, et al. A comparative study of autoregressive neural network hybrids. Neural Networks 2005; 18: 781–789.

[52] M. D. Philemon, Z. Ismail, and J. Dare, "A review of epidemic forecasting using artificial neural networks," International Journal of Epidemiologic Research, vol. 6, no. 3, pp. 132–143, 2019.

[53] Aslanargun A, et al. Comparison of ARIMA, neural networks and hybrid models in time series: tourist arrival forecasting. Journal of Statistical Computation and Simulation 2007; 77: 29–53.

[54] Jain A, et al. Hybrid neural network models for hydrologic time series forecasting. Applied Soft Computing 2007; 7: 585–592.

[55] Yu L, et al. Application of a new hybrid model with seasonal auto-regressive integrated moving average (ARIMA) and nonlinear auto-regressive neural network (NARNN) in forecasting incidence cases of HFMD in Shenzhen, China. PLoS ONE 2014; 9: e98241.

[56] Zhang GP. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 2003; 50: 159–175.

[57] Wang, K. W., et al. "Hybrid methodology for tuberculosis incidence time-series forecasting based on ARIMA and a NAR neural network." Epidemiology & Infection 145.6 (2017): 1118-1129.

[58] Benmouiza K, Cheknane A. Small-scale solar radiation forecasting using ARMA and nonlinear autoregressive neural network models. Theoretical & Applied Climatology 2016; 124: 945–958.

[59] Box GEP, Jenkins G. Time Series Analysis, Forecasting and Control. San Francisco, CA: Holden-Day, 1970.

[60] C.W.J. Granger, Combining forecasts—Twenty years later, J. Forecasting 8 (1989) 167–173.

[61] Dickey DA, Fuller WA. Likelihood ratio statistics for autoregressive time series with a unit root. Econometrica, 1981;1057-1072.

[62] Levenberg K. A method for the solution of certain problems in least squares. Q Appl Math 1944; 5:164–8.

[63] MacKay DJC. Bayesian interpolation. Neural Comput 1992;4(3):415–47.

[64] Moller MF. A scaled conjugate gradient algorithm for fast supervised learning. Neural Networks 1993; 4:525–33.

[65] Beale MH, et al. Demuth. Neural Network Toolbox™ Getting Started Guide. The MathWorks Inc. 2012.

Figures

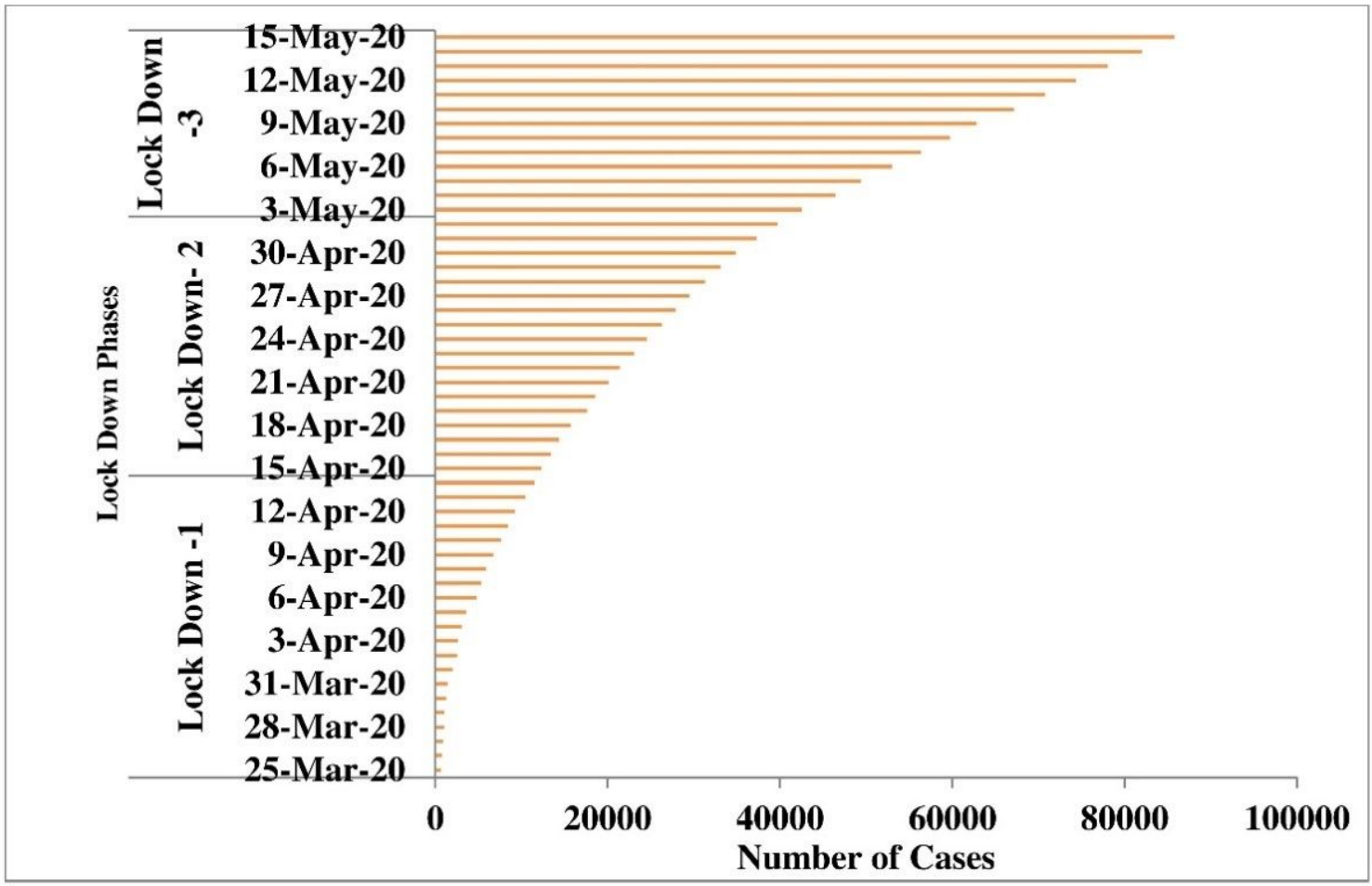

Figure 1

Confirmed cases of COVID-19 in India during Lockdown

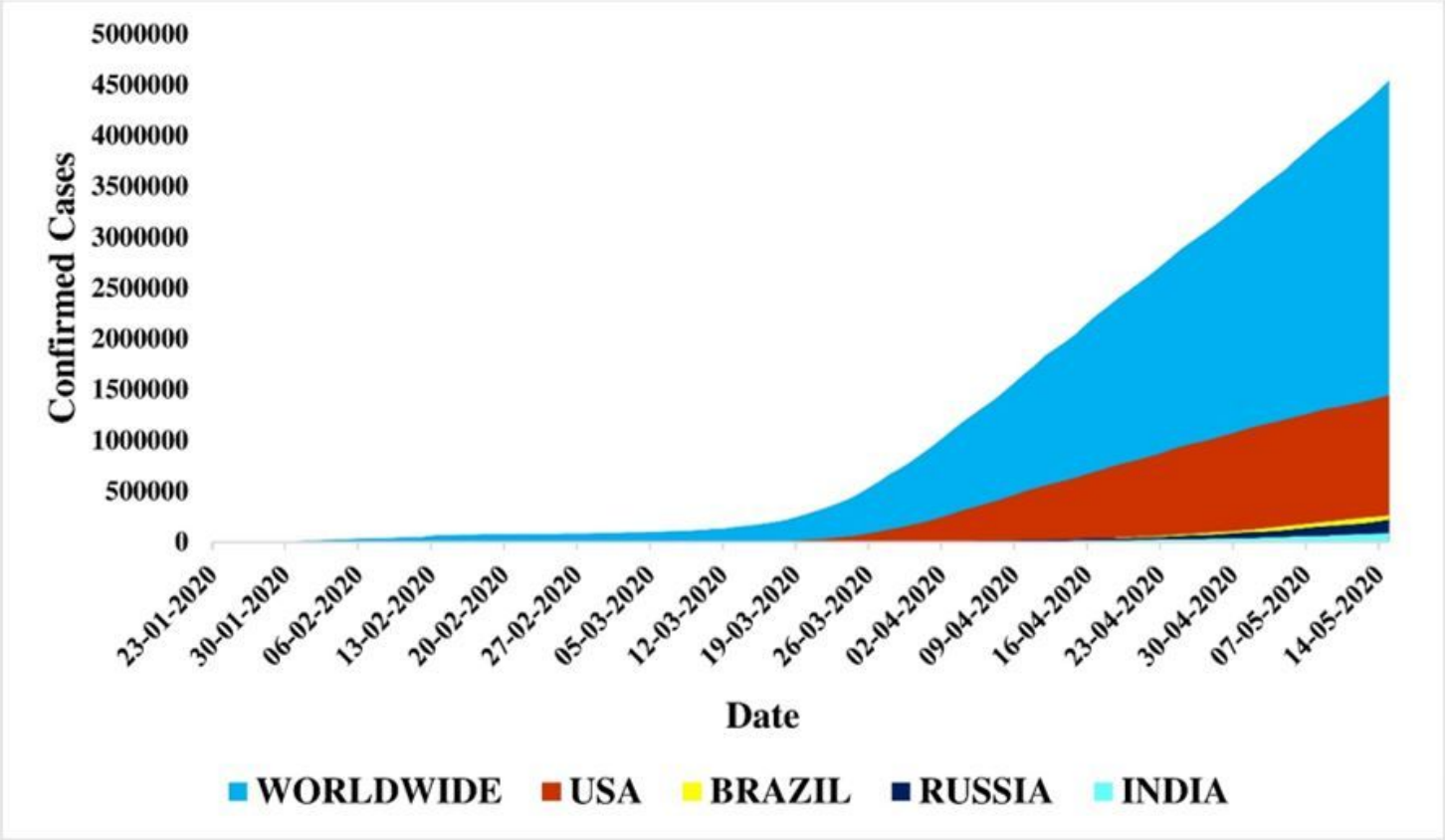

Figure 2

Total Confirmed cases of COVID-19 Worldwide from Jan 22 to May 15, 2020[1]

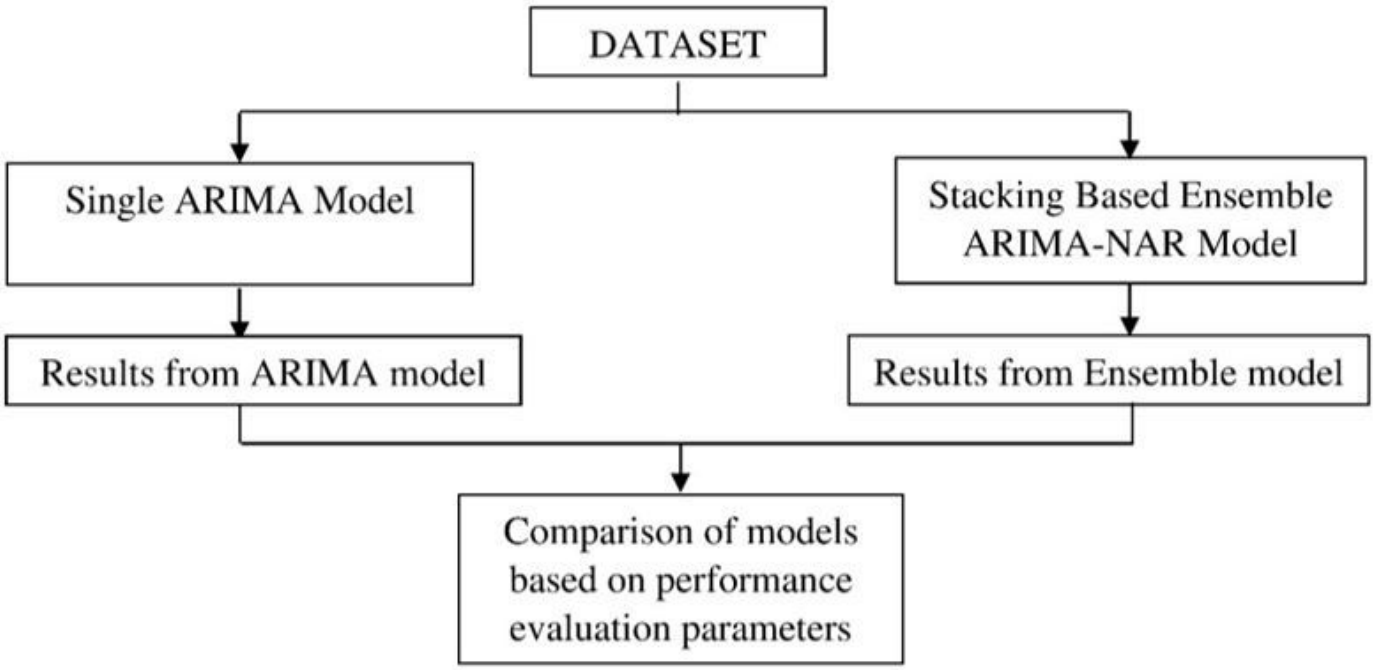

Figure 3

Flow chart indicating various steps involved for forecasting.

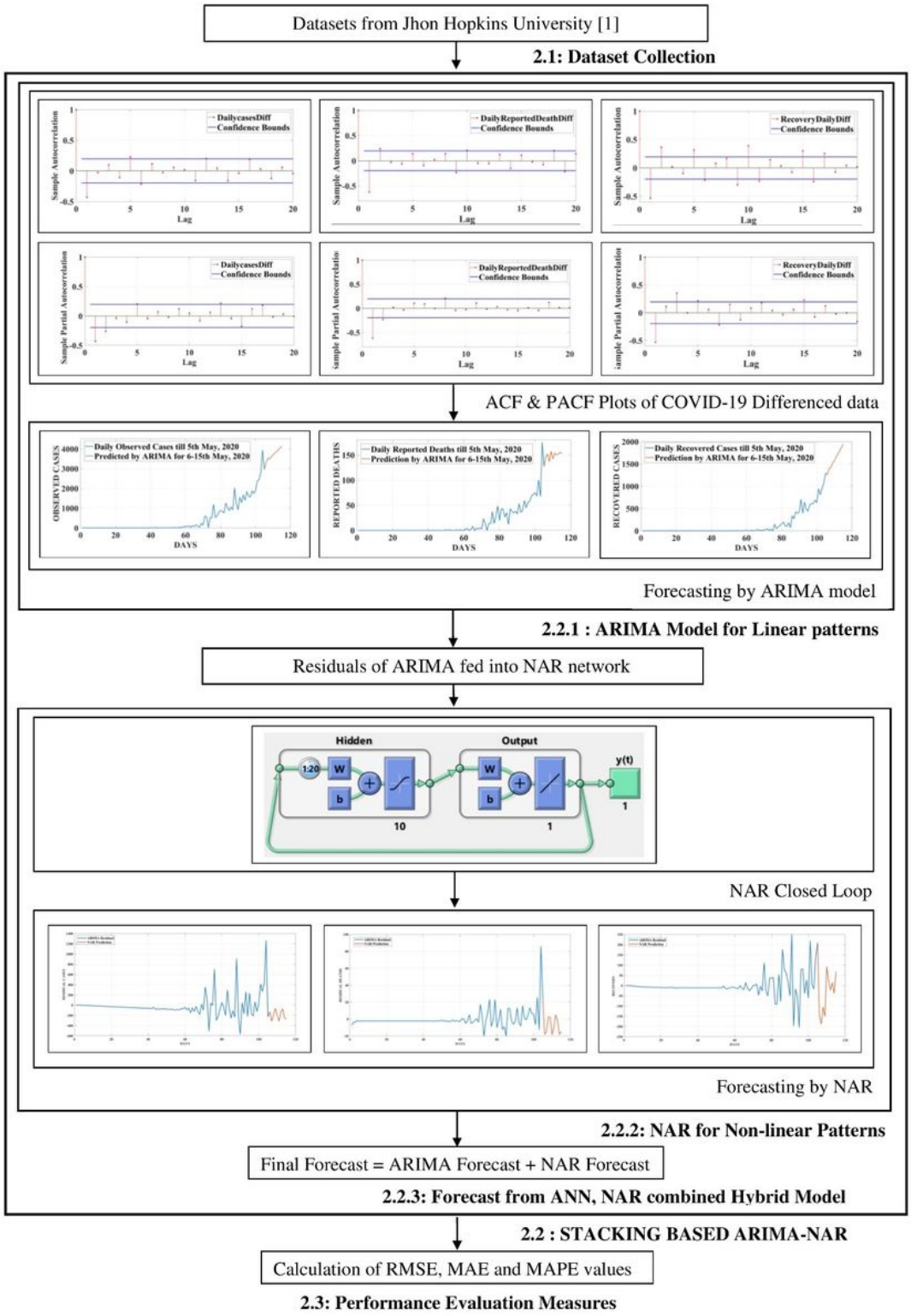

Figure 4

Pictorial description of the stack based ensemble ARIMA model

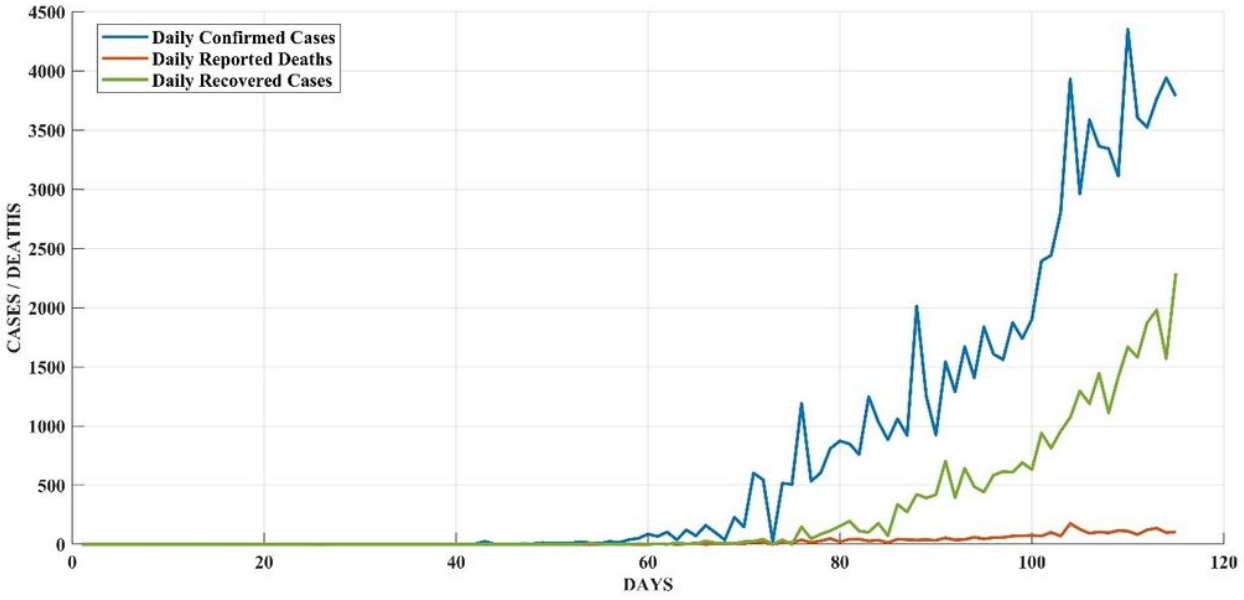

Figure 5

Daily observed cases, reported deaths and recovered cases of COVID-19 in India till May 15, 2020

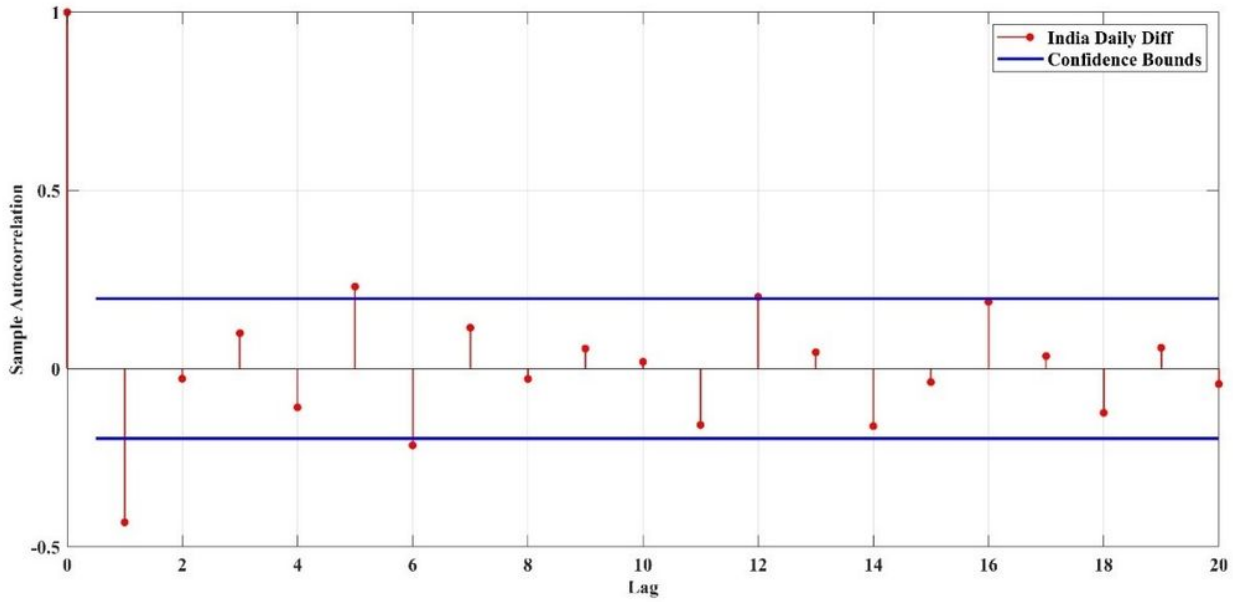

Figure 6

Auto-correlation Function of India_Daily_1stDifference

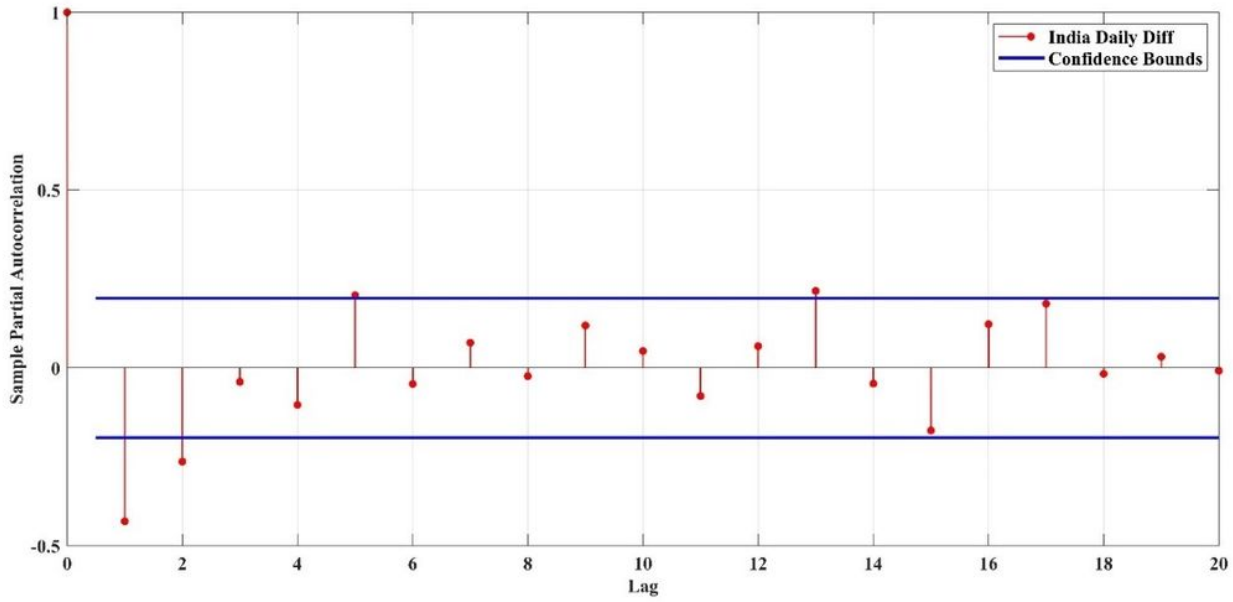

Figure 7

Partial Auto-correlation Function of India_Daily_1st Difference

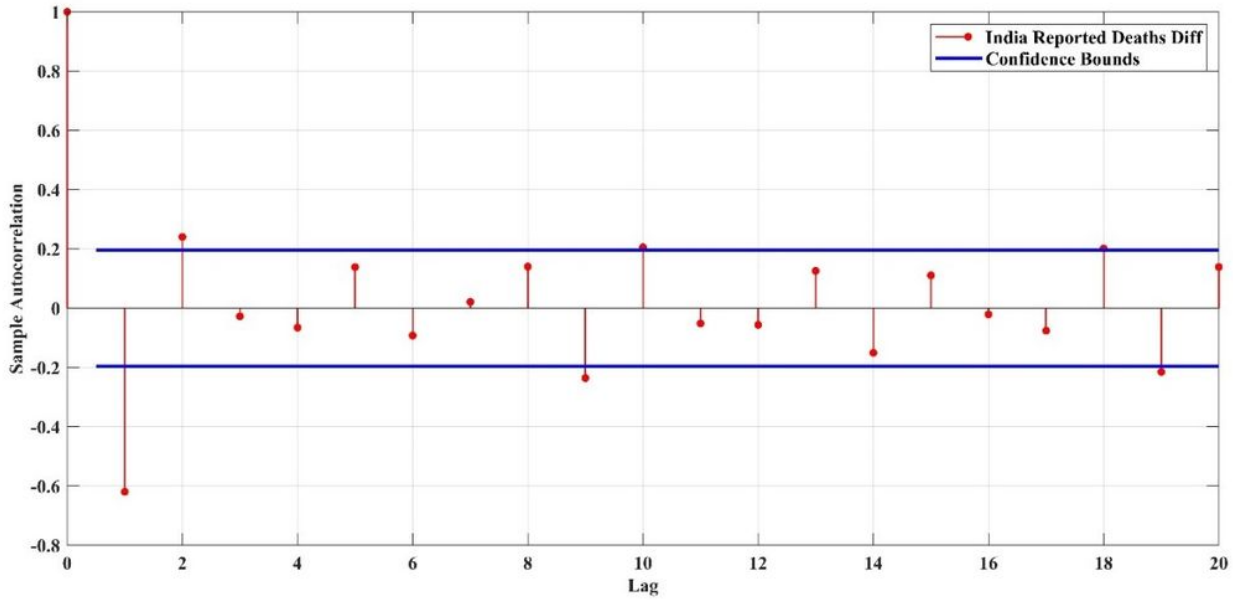

Figure 8

Auto-correlation Function of India_Deaths_1st Difference

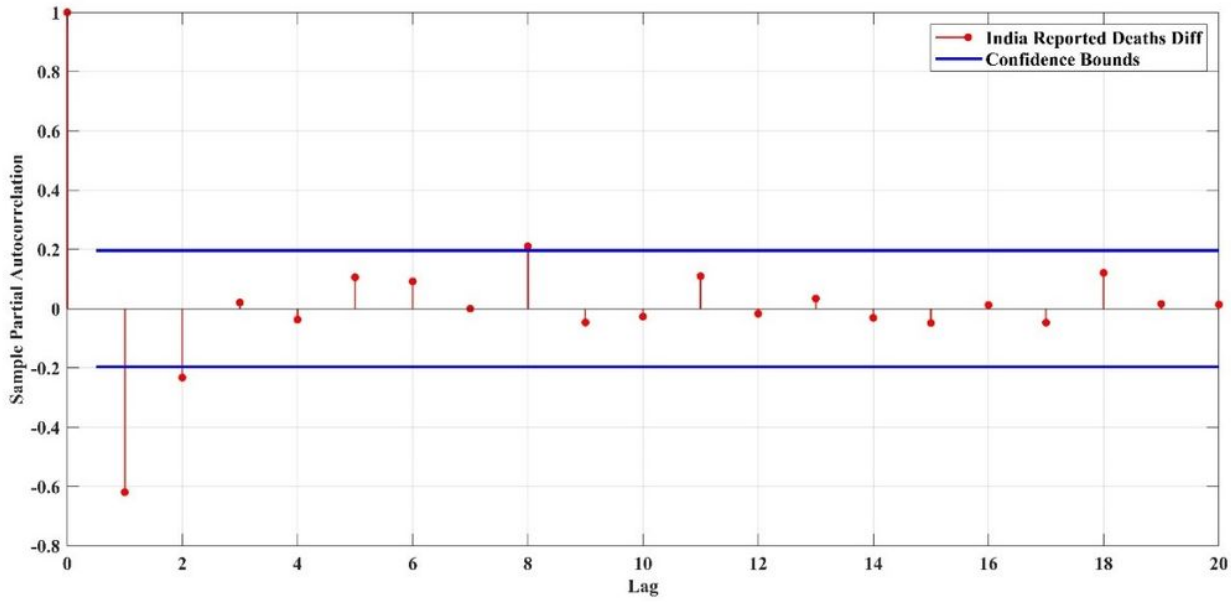

Figure 9

Partial Auto-correlation Function of India_Deaths_1st Difference

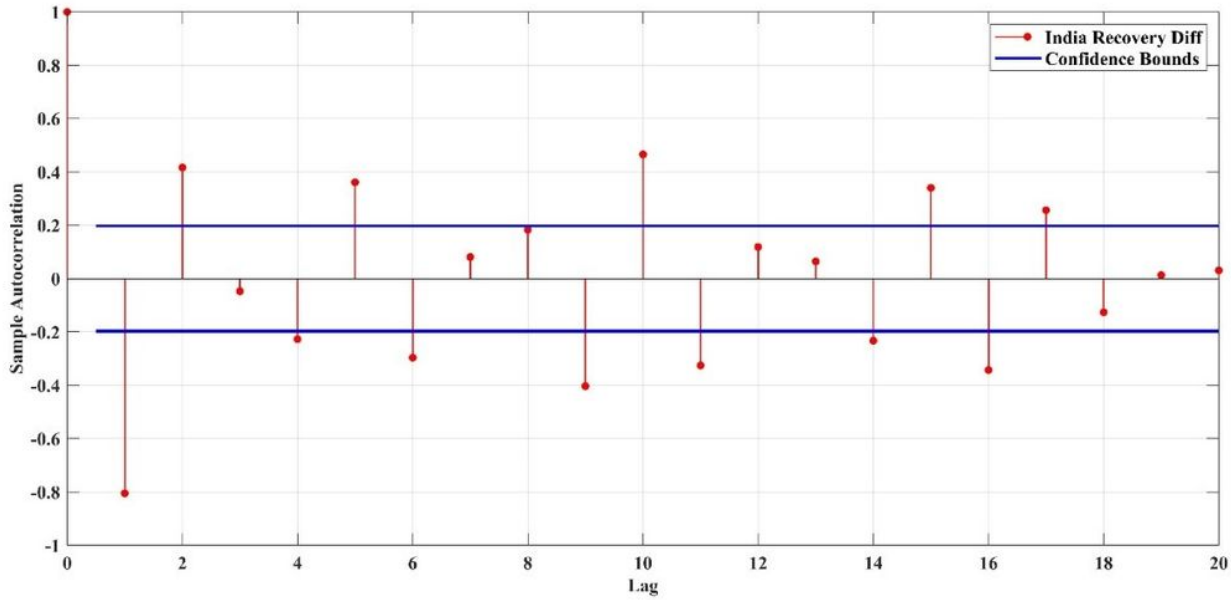

Figure 10

Auto-correlation Function of India_Recovery_1st Difference

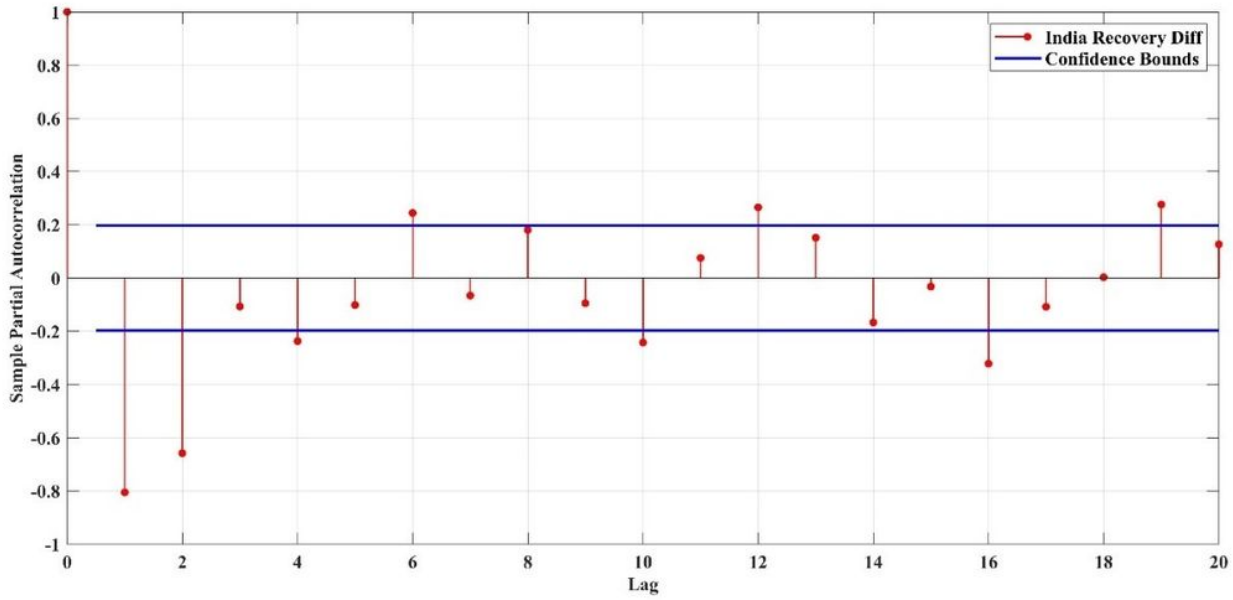

Figure 11

Partial Auto-correlation Function of India_Recovery_1st Difference

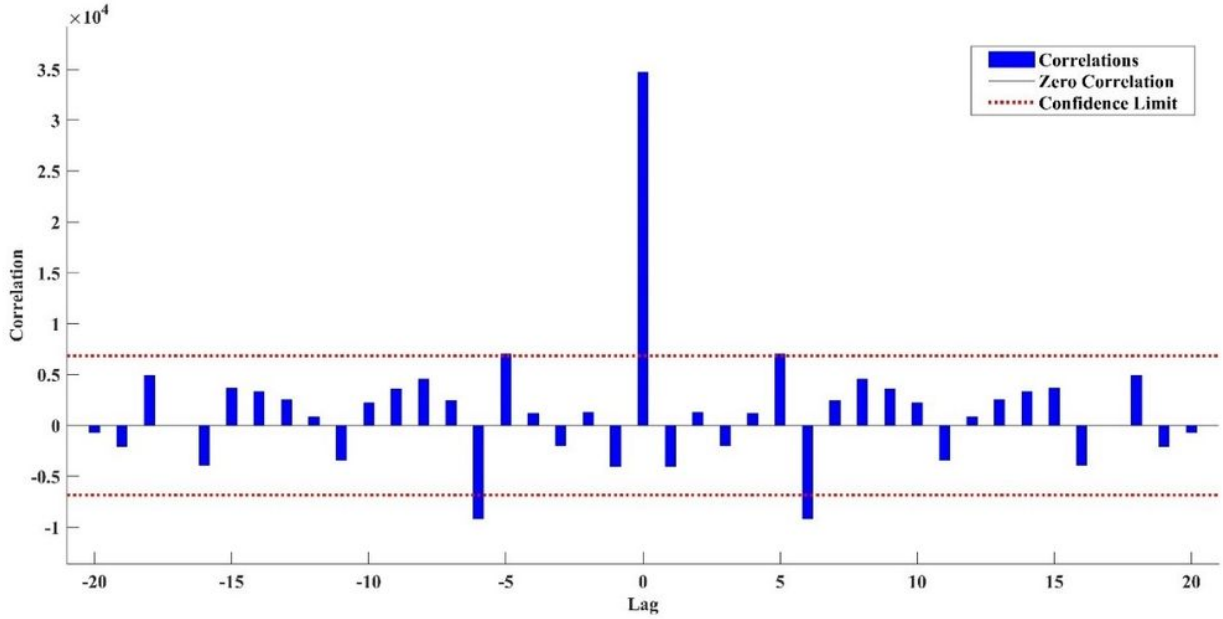

Figure 12

Error auto-correlation plot for NAR-Residual of Daily Confirmed Cases

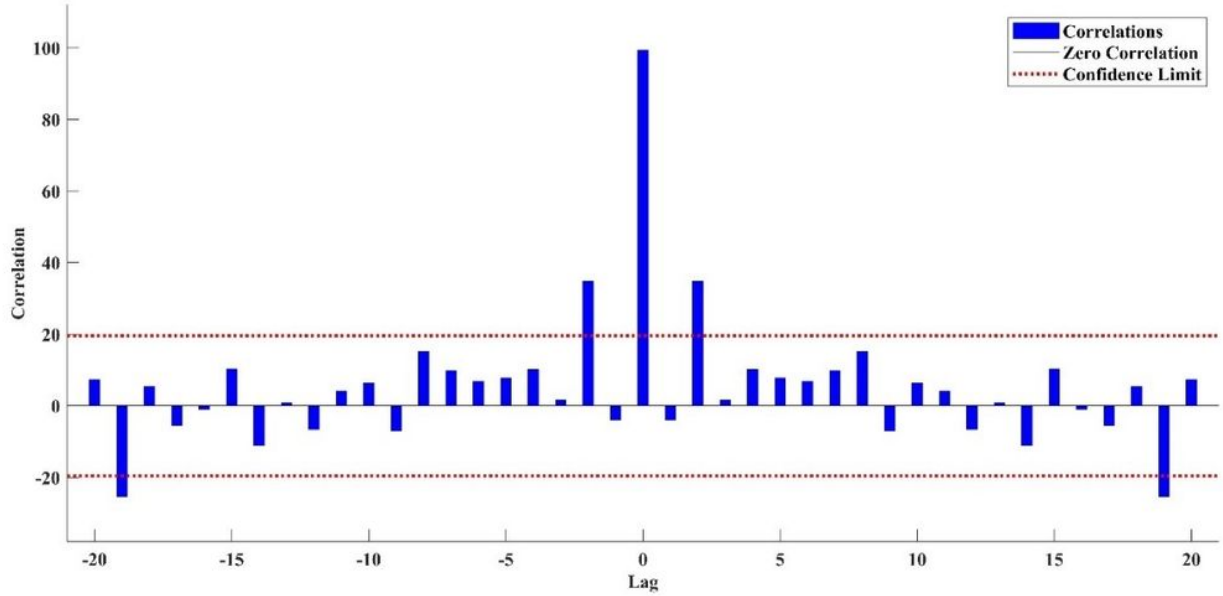

Figure 13

Error auto-correlation plot for NAR-Residual of Daily Reported Deaths

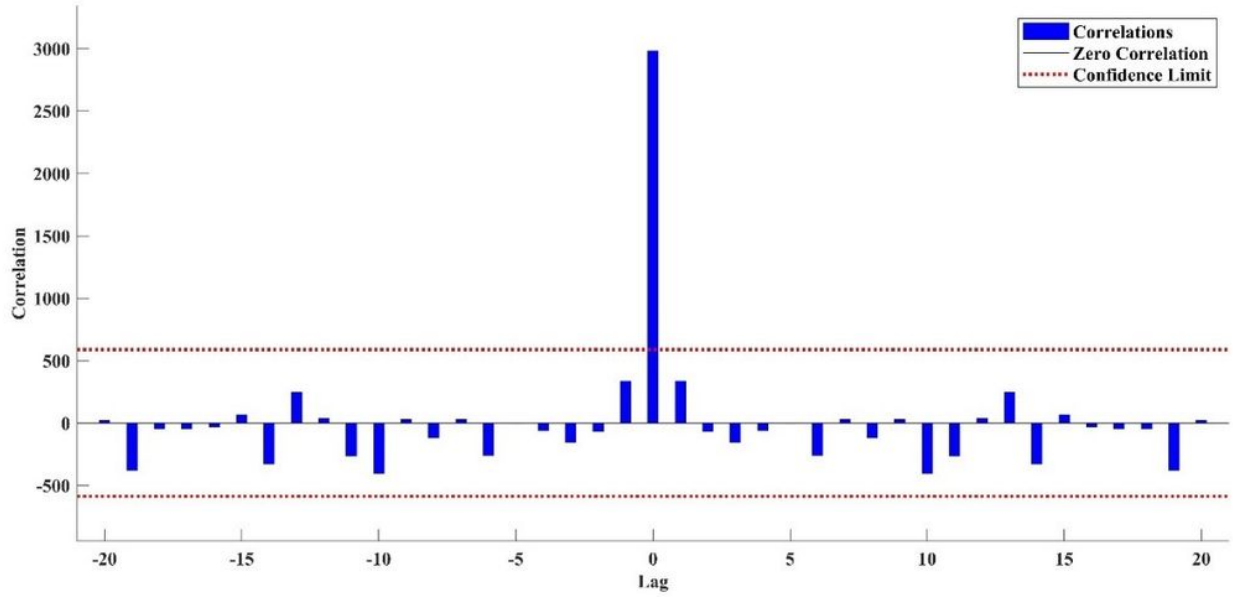

Figure 14

Error auto-correlation plot for NAR-Residual of Daily Recovered Cases

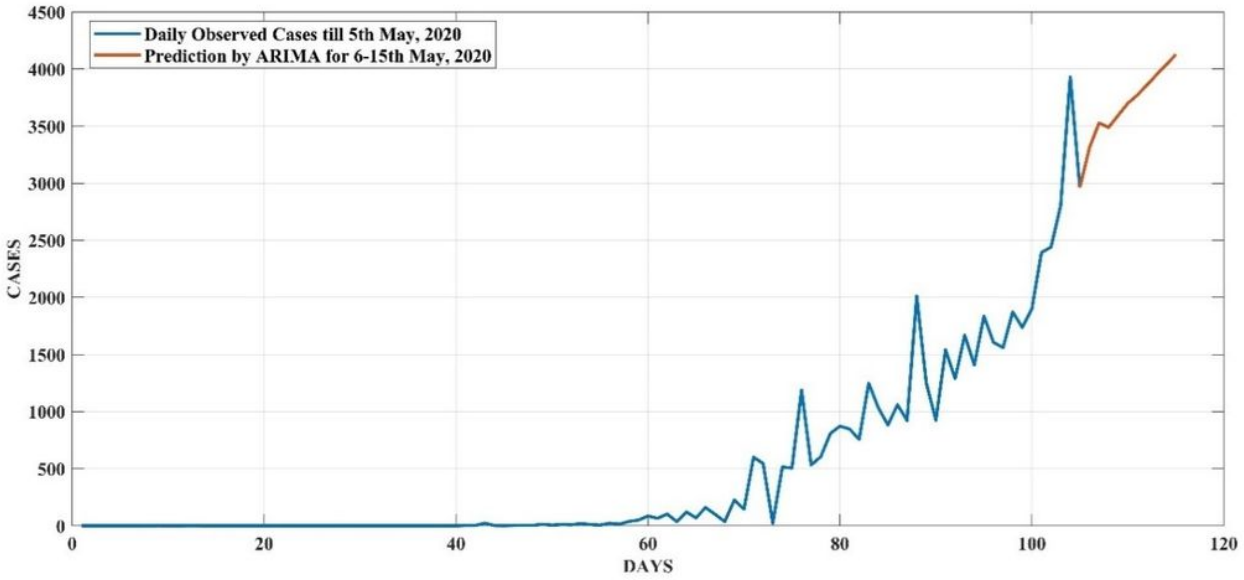

Figure 15

Prediction by ARIMA model for Daily observed cases in India between 6th and 15th May, 2020

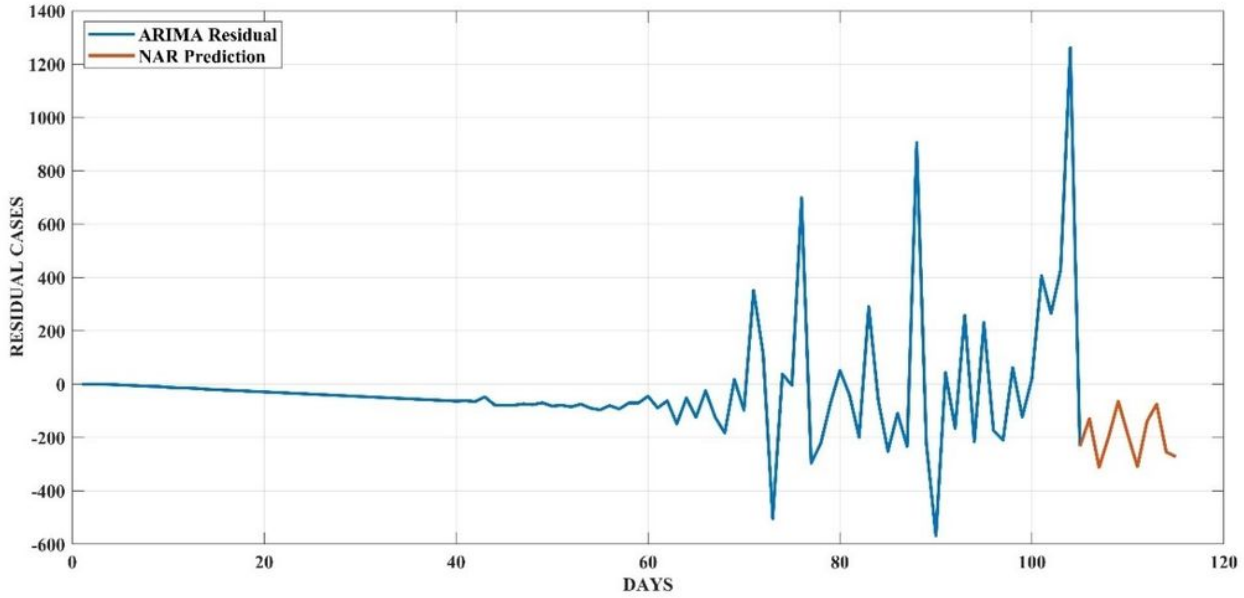

Figure 16

Prediction of residual error by NAR model for daily observed cases in India between 6th and 15th May, 2020

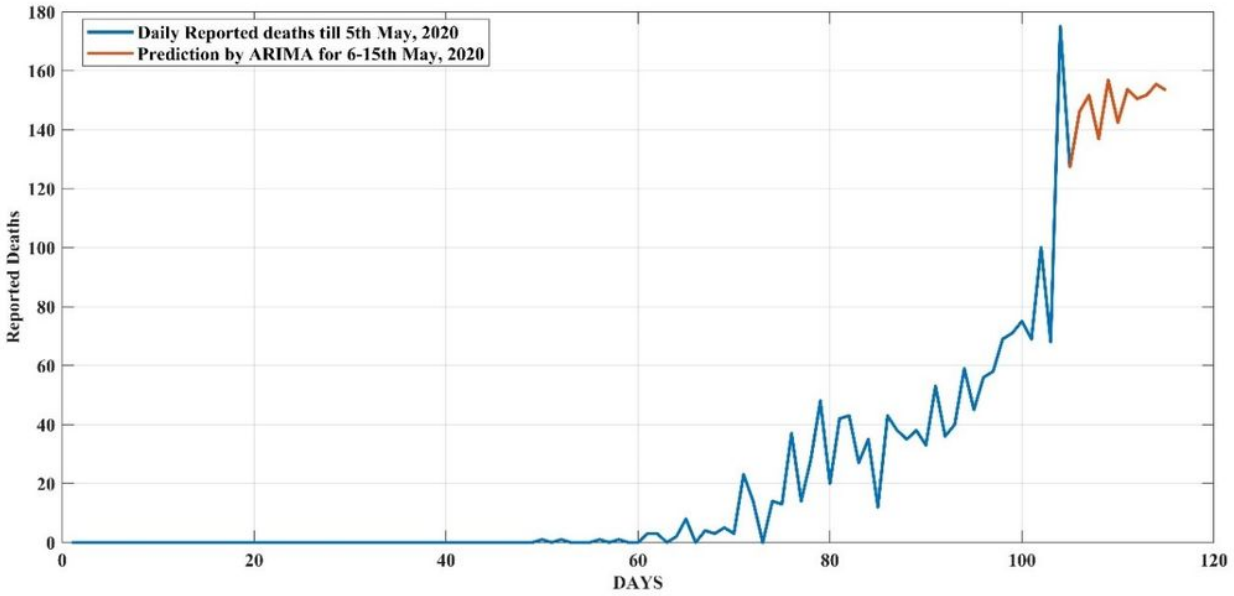

Figure 17

Prediction by ARIMAmodel for Daily Reported deathsin India between 6th and 15th May, 2020

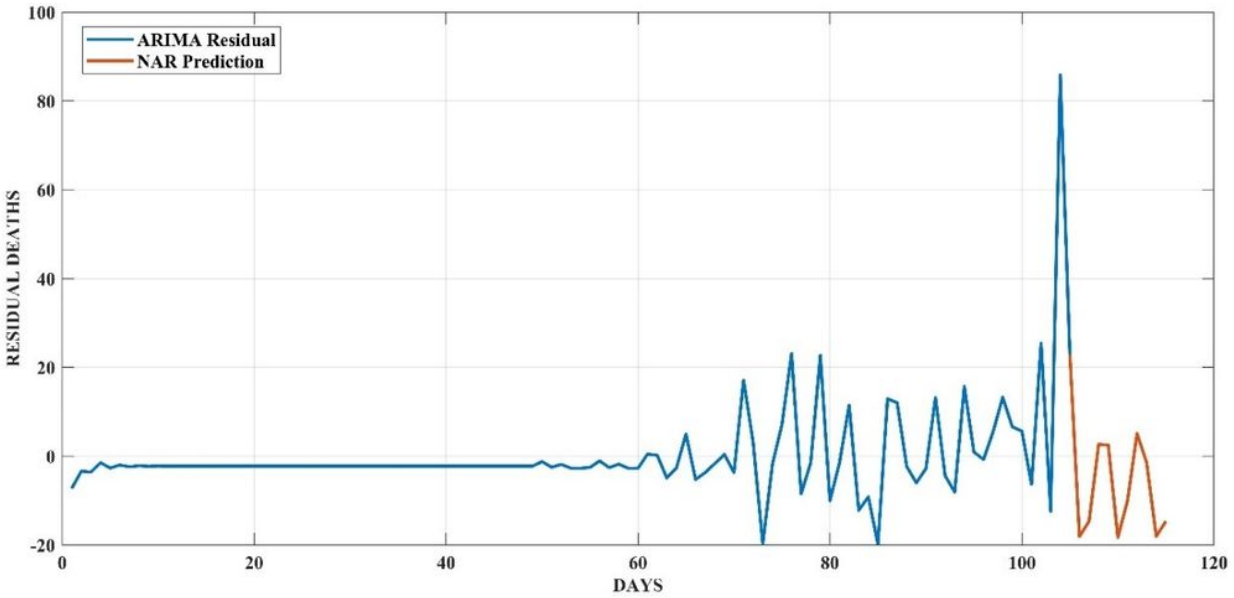

Figure 18

Prediction of residual error by NAR model for daily reported deaths in India between 6th and 15th May, 2020

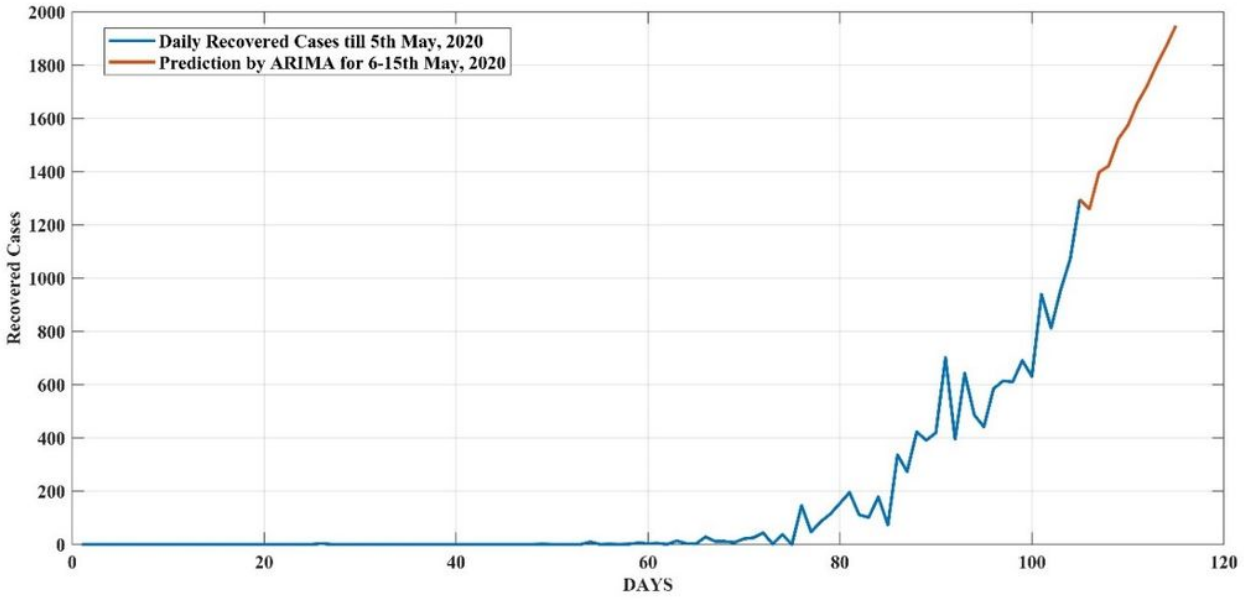

Figure 19

Prediction by ARIMA model for Daily recovered cases in India between 6th and 15th May, 2020

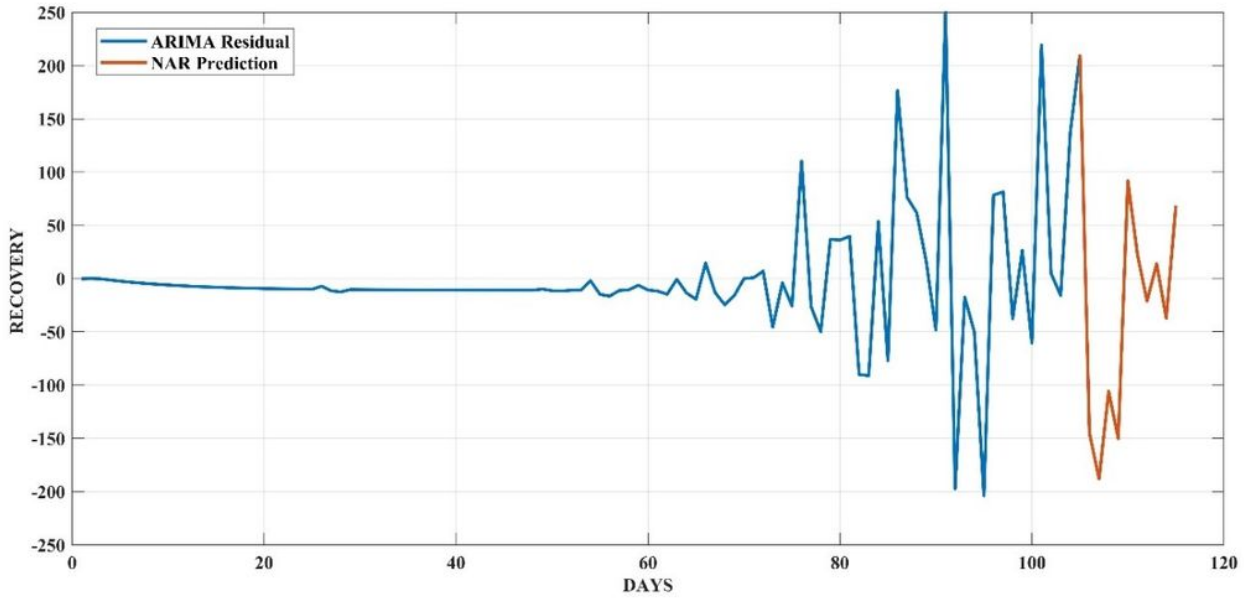

Figure 20

Prediction of residual error by NAR model for daily recovered cases in India between 6th and 15th May, 2020