Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Journal of Biomedical Informatics 121 (2021) 103887

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Implementation of stacking based ARIMA model for prediction of Covid-19 cases in India

Aman Swaraj [f,1,*], Karan Verma [a], Arshpreet Kaur [b], Ghanshyam Singh [c], Ashok Kumar [d], Leandro Melo de Sales [e]

[a] National Institute of Technology, Delhi, India
[b] DIT University, Dehradun, India
[c] Malaviya National Institute of Technology Jaipur, India
[d] Government Mahila Engineering College, Ajmer, India
[e] Universidade Federal De Alagoas-UFAL, Brazil
[f] Indian Institute of Technology, Roorkee, India

ARTICLE INFO

Keywords:
Hybrid model
Forecasting
COVID-19
ARIMA
NAR

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.

Abbreviations: 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; 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; 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.

* Corresponding author at: Main Campus, Indian Institute of Technology, Roorkee, 247667 Uttarakhand, India.
E-mail addresses: amanswaraj007@gmail.com (A. Swaraj), karanverma@nitdelhi.ac.in (K. Verma), arshpreet.kaur@dituniversity.edu (A. Kaur), gsingh.ece@mnit.ac.in (G. Singh), kumarashoksaini@gmail.com (A. Kumar), leandro@ic.ufal (L. Melo de Sales).
[1] S-488, Anukampa Bhawan, Near Yadav Dairy, Solanipuram, Roorkee, 247667 Uttarakhand, India.

https://doi.org/10.1016/j.jbi.2021.103887
Received 5 August 2020; Received in revised form 1 June 2021; Accepted 8 August 2021
Available online 15 August 2021

Table 1
Depiction of lockdown phases.

Lock down phases	Dates	Number of cases	Days	Increase percentage
Phase 0	22/01/2020, 24/03/2020	2872	58	–
Phase 1	25/03/2020–14/04/2020	10,951	21	281.3%
Phase 2	15/03/2020 – 02/05/2020	31,118	19	184.16%
Phase 3	03/05/2020–17/05/2020	53,193	12	70.93%

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.

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.

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. 1). 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.

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].

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 [28–43]. Some of the relevant work is presented in Table 2.

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

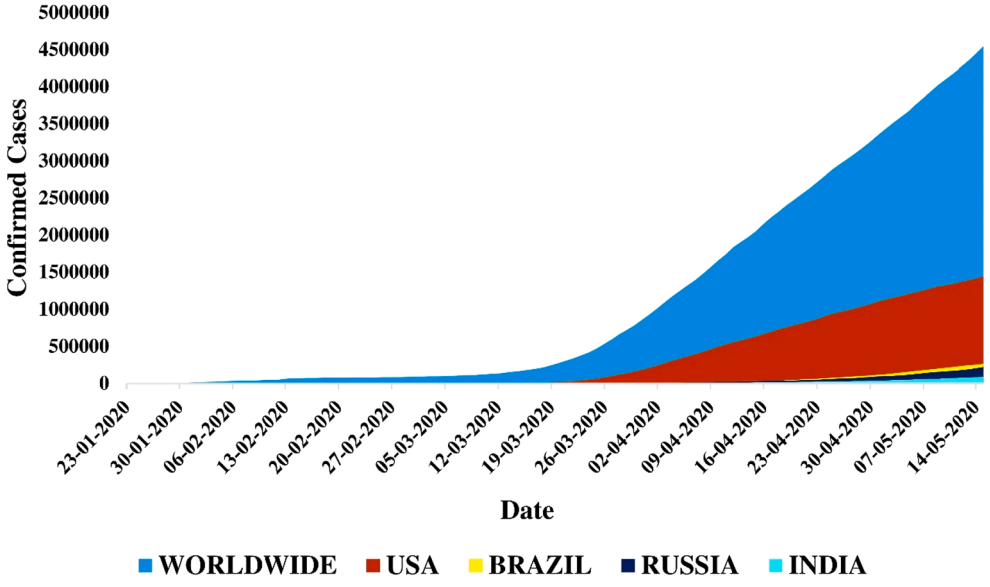


Fig. 1. Total Confirmed cases of COVID-19 Worldwide from Jan 22 to May 15, 2020 [1].

Table 2
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	

(ABC – Artificial Bee Colony; KNN – K-Nearest Neighbors; Support Vector Regression (SVR); ANFIS – Adaptive Neuro-Fuzzy Inference System; PSO – Particle Swarm Optimization; GA – Genetic Algorithm; FPA – Flower Pollination Algorithm; FPASSA – Flower pollination algorithm Salp Swarm Algorithm; ARIMA – Auto -Regressive Integrated Moving Average; DNN – Deep Neural Network; LSTM – Long short-term memory; PR- Polynomial Regression; ANN – Artificial Neural Networks).

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

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. 2. 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.

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 for three phases, between may 6–15, July 21–30 and Aug 1–10. The starting point however is fixed at 22nd January.

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. Fig. 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:

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

Here, A_t is the actual observed value at time t and E_t is random error. $\phi_i (i = 1, 2, \cdots, a)$ and $\theta_j (j = 0, 1, 2, \cdots, c)$ 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. Snd 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:

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

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:

$$\widehat{\boldsymbol{Z}}(\boldsymbol{t}) = \boldsymbol{fx}(\boldsymbol{Z}(\boldsymbol{t}-1) + \boldsymbol{Z}(\boldsymbol{t}-2) + \cdots + \boldsymbol{Z}(\boldsymbol{t}-\boldsymbol{n}))) \quad (3)$$

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.

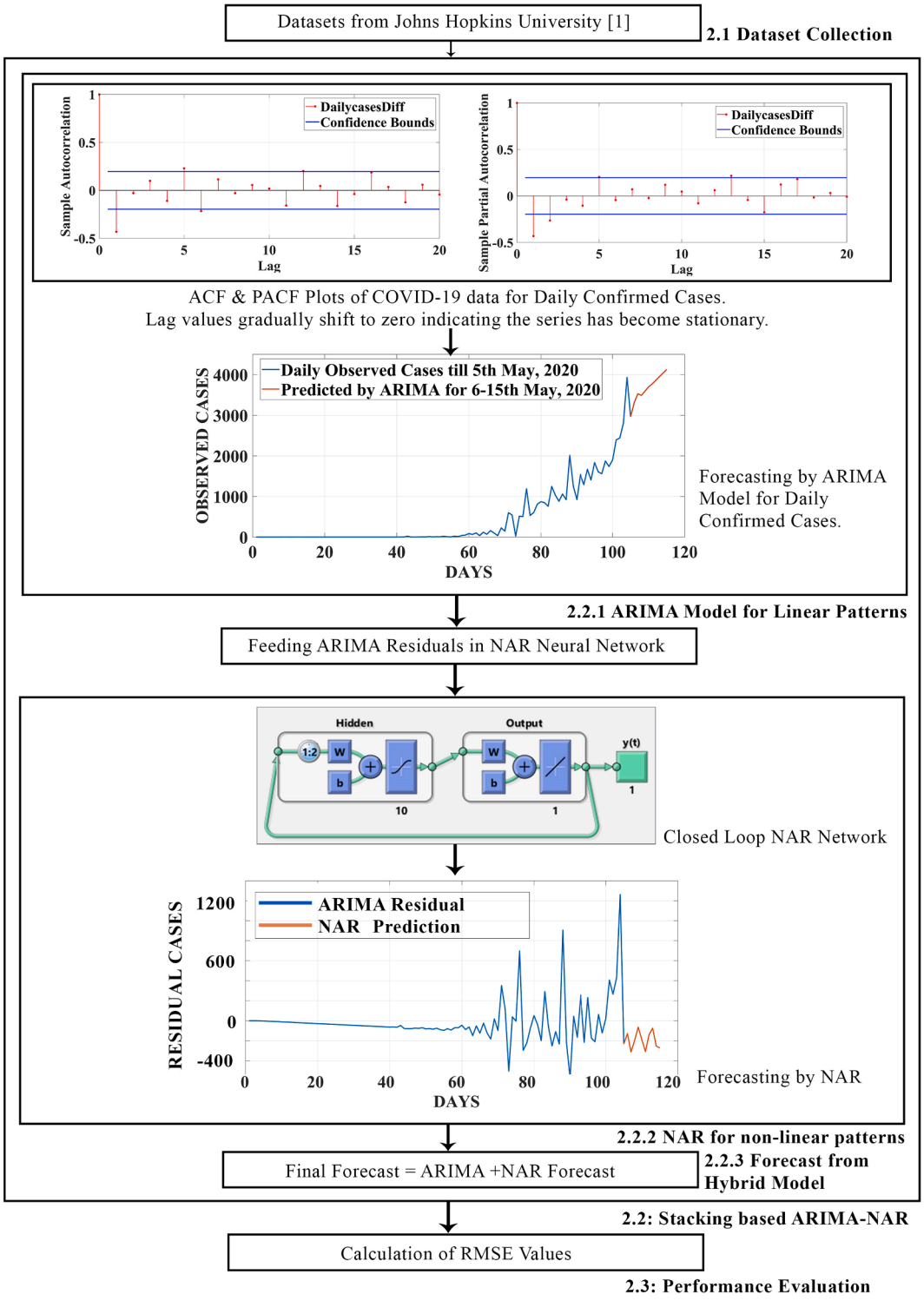


Fig. 2. Pictorial description of the stack based ensemble ARIMA model.

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

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:

$$\boldsymbol{Z}_t = \boldsymbol{L}_t + \boldsymbol{N}_t, \quad (5)$$

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

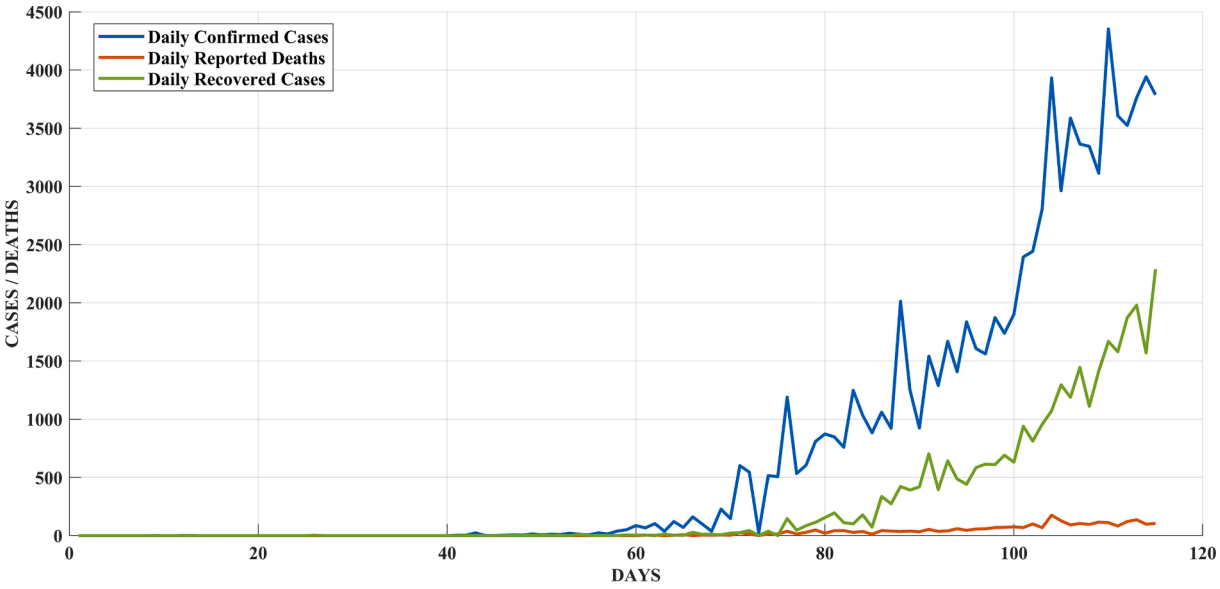


Fig. 3. Daily observed cases, reported deaths and recovered cases of COVID-19 in India till May 15, 2020.

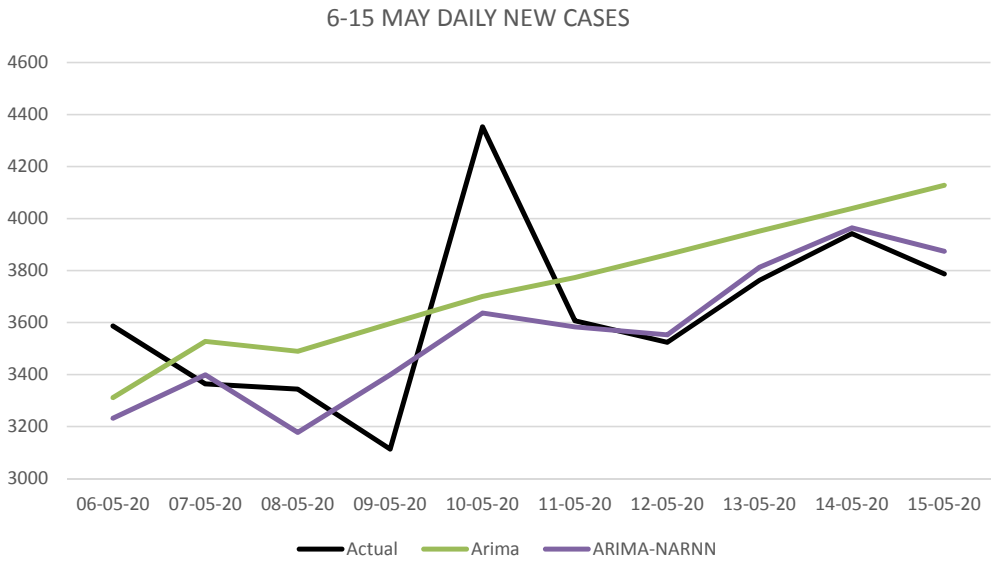


Fig. 4. Prediction by ARIMA, ARIMA-NAR Model for daily new cases of COVID-19 in India between May 6–15, 2020.

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,

$$\boldsymbol{R}_t = \boldsymbol{Z}_t - \widehat{\boldsymbol{L}}_t, \tag{6}$$

where, $\widehat{L}_t$ 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:

$$\boldsymbol{R}_t = \boldsymbol{fx}(\boldsymbol{R}_{t-1}, \boldsymbol{R}_{t-2}, \cdots, \boldsymbol{R}_{t-n}) + \epsilon_t, \tag{7}$$

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:

$$\widehat{\boldsymbol{Z}}_t = \widehat{\boldsymbol{L}}_t + \widehat{\boldsymbol{N}}_t, \tag{8}$$

where, $\widehat{Z}_t$ denotes the final predicted values at time t, and Eq. (7) is represented as $\widehat{N}_t$, 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.

3. Constructing the hybrid model in MATLAB

Data is first divided into training, testing and validation randomly on multiple iterations. Several weight optimising algorithms are then used for adjusting the weight values, and the 'Neural Net Time Series

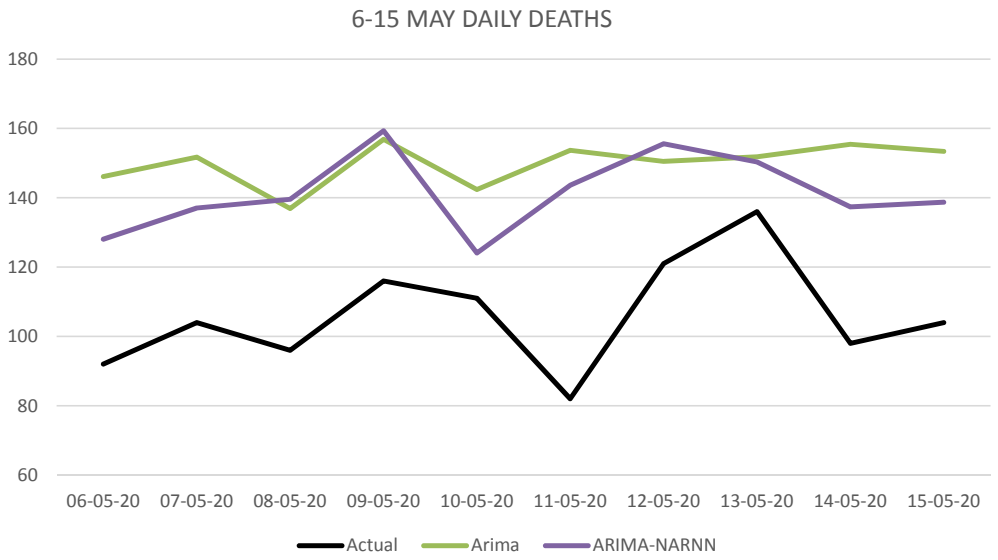


Fig. 5. Prediction by ARIMA, ARIMA-NAR Model for daily new reports of death due to COVID-19 in India between May 6–15, 2020.

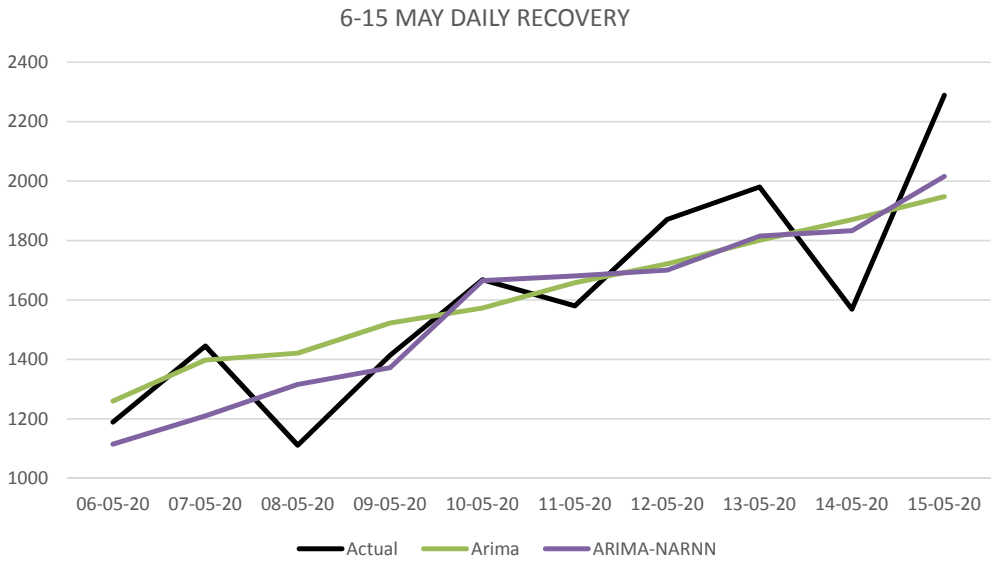


Fig. 6. Prediction by ARIMA, ARIMA-NAR Model for daily new cases of recovery from COVID-19 in India between May 6–15, 2020.

Table 3a
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	7.8%
Hybrid arima	**275.9648**	**176.9298**	**4.7%**

Table 3b
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	44.02%
Hybrid arima	**37.79482**	**35.3597**	**35.32%**

Table 3c
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	10.66%
Hybrid Arima	**177.6032**	**153.3469**	**9.67%**

Toolbox' in MATLAB provides three sets of such algorithms, namely Levemberg–Marquardt [61], Bayesian Regularization [62] and scaled conjugate gradient [63]. Low MSE and higher R values account for selection the optimum NAR model. The error autocorrelation plot is also used for verifying the adequacy of the model. After the training is finished, all the synaptic weights are saved, and the model is ready for prediction.

3.1. Performance evaluation measures

In general, the performance of any forecasting model is determined

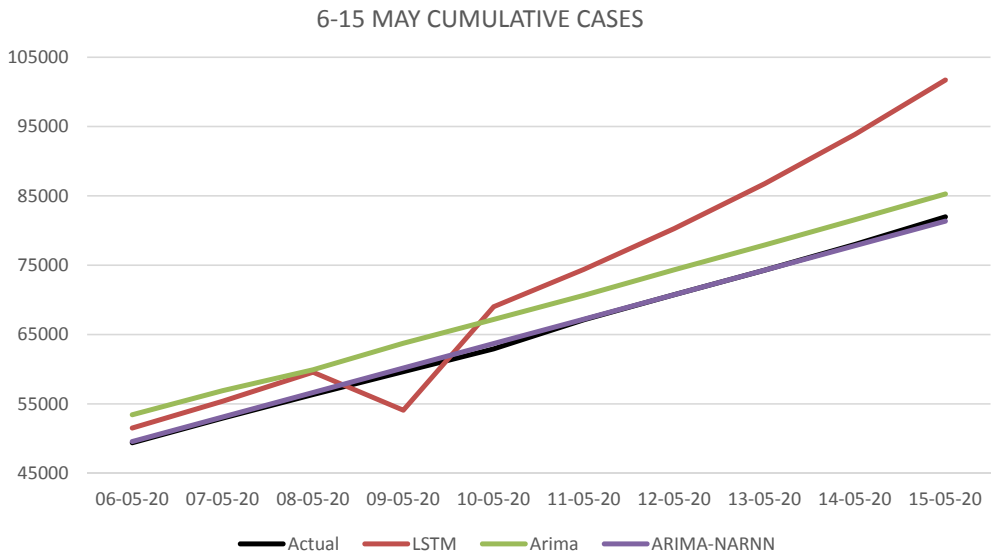


Fig. 7. Prediction by ARIMA, ARIMA-NAR and LSTM Model for cumulative new cases of COVID-19 in India between May 6–15, 2020.

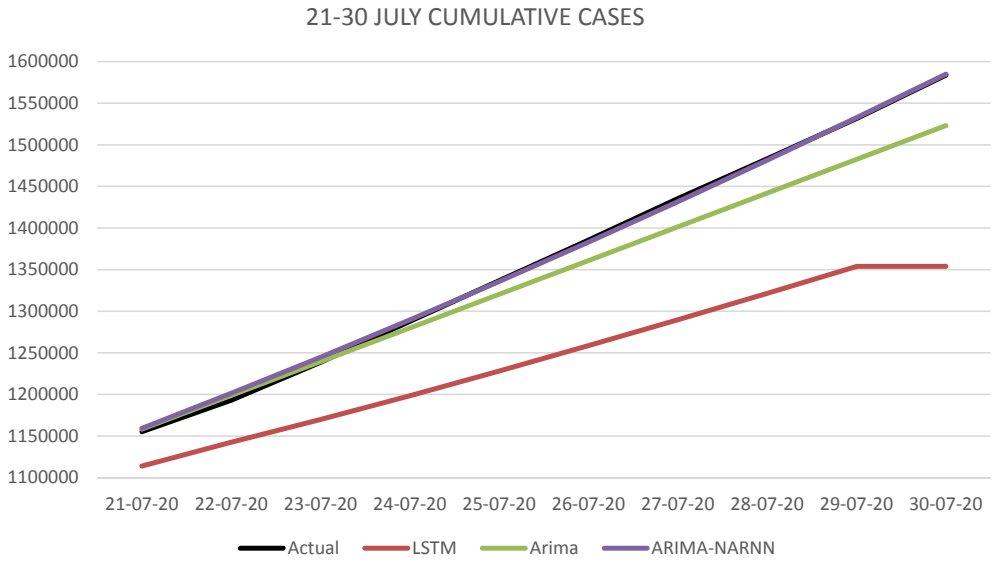


Fig. 8. Prediction by ARIMA, ARIMA-NAR and LSTM Model for cumulative new cases of COVID-19 in India between July 21–30, 2020.

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.

$$\boldsymbol{RMSE} = \sqrt{\frac{1}{\boldsymbol{n}}\sum_{t=1}^{n}\left(\boldsymbol{Z}_t - \widehat{\boldsymbol{Z}}_t\right)^2} \tag{9}$$

$$\boldsymbol{MAE} = \frac{1}{\boldsymbol{n}}\sum_{t=1}^{n}\left|\boldsymbol{Z}_t - \widehat{\boldsymbol{Z}}_t\right| \tag{10}$$

$$\boldsymbol{MAPE} = \frac{1}{\boldsymbol{n}}\sum_{t=1}^{n}\frac{\left|\boldsymbol{Z}_t - \widehat{\boldsymbol{Z}}_t\right|}{\boldsymbol{Z}_t} \tag{11}$$

4. 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. Fig. 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 and additionally for 21–30 July and 1–10 Aug for cumulative cases in India. We also compare the results with LSTM and SIR model.

The final forecasting is done by combining the separate prediction values of ARIMA and NAR models. Figs. 4–6 respectively show the prediction of future cases by the ARIMA and NAR neural network for daily observed cases, reported deaths, and daily recovered cases between May 6–15, 2020. 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 (Tables 3a–3c). Figs. 7–9 further draw a comparision between three different models, ARIMA, Hybrid ARIMA and LSTM for cumulative cases of covid-19 in India for three different phases, respectively 6–15 May, 21–30 July, 1–10 Aug. Additionally, we also draw comparision with the compartmental model,

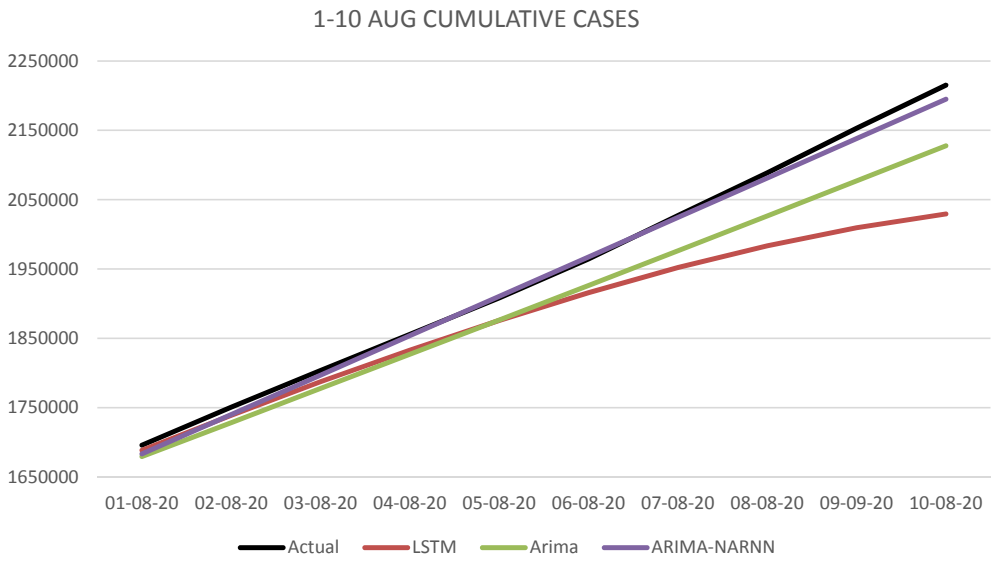


Fig. 9. Prediction by ARIMA, ARIMA-NAR and LSTM Model for cumulative new cases of COVID-19 in India between Aug 1–10, 2020.

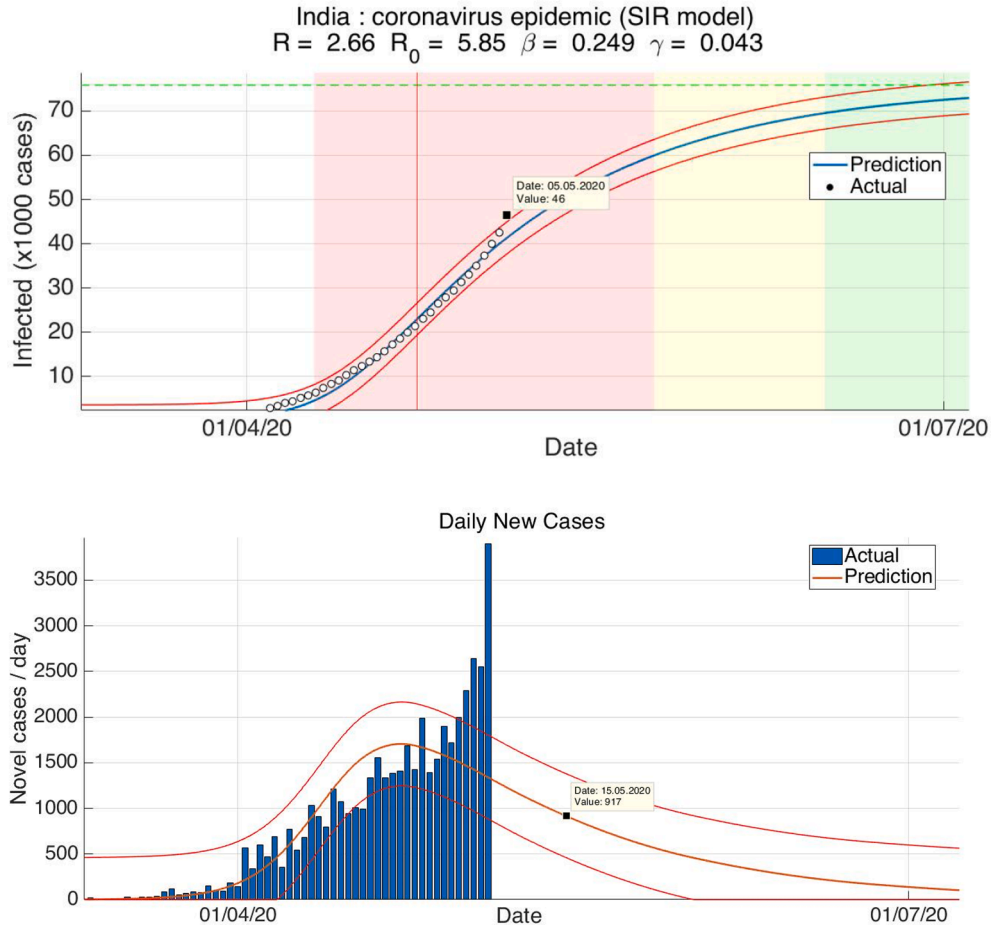


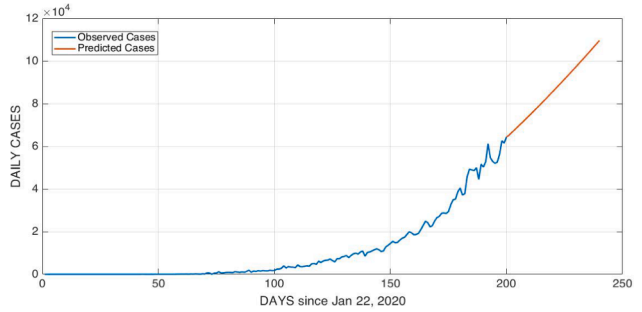
Fig. 10. Predictions using the SIR model. Top panel with white, red, yellow and green regions indicate initial exponential growth, fast growth (with positive and negative phase separated by red vertical line), asymptotic slow growth and curve flattening, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

SIR in Fig. 10 and Table 4. Finally, we also present long term forecast of covid-19 cases with the hybrid model (Fig. 11) and Table 5.

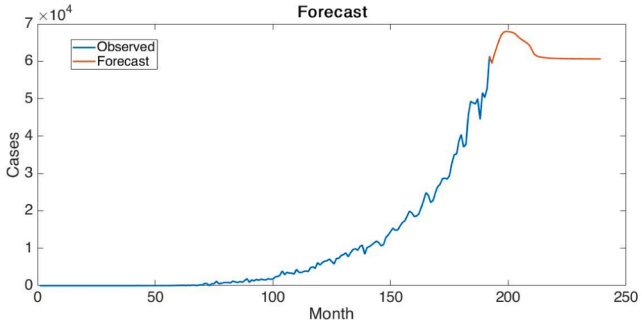
As seen in Tables 3a–3c, hybrid ARIMA's performance provide more adequate results. The RMSE, MAE and MAPE value of the hybrid combination for daily observed cases are 275.9648 (16.23% reduction), 176.9298 (37.89% reduction), 4.7% (39.53% reduction). Regarding daily reported deaths, cases of recovery and cumulative confirmed cases similar results were found with reduced error percentages. Further, it is

Table 4
Accuracy comparison of SIR model and Hybrid Arima model for daily new cases in India between 6th and 15th May 2020.

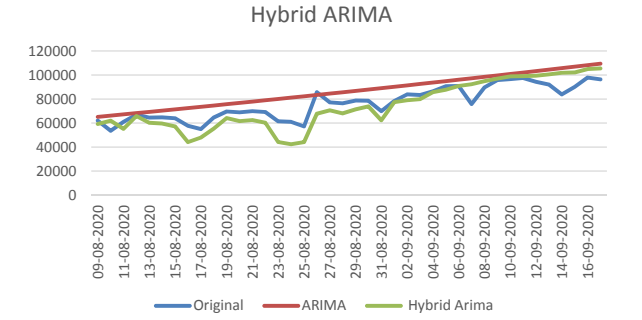
	SIR model	Hybrid model
RMSE	**2499.233**	**275.9648**



(a)



(b)



(c)

Fig. 11. Forecast for a duration of 40 days using (a) ARIMA; (b) LSTM; (c) Hybrid ARIMA.

Table 5
Accuracy comparison of ARIMA model and Hybrid Arima model for a duration of 40 days.

	ARIMA model	Hybrid model
RMSE	**11759.72517**	**8908.786344**

evident from Figs. 4–9 that Hybrid ARIMA has consistently performed better on all occasions.

Prediction with SIR model-

The well known compartmental model, Susceptible-Infectious-Recovered (SIR) model deals with the number of susceptibles 'S'; number of infectious 'I'; and the number of recovered or deceased individuals 'R'. Details of the implementation including selection of R_0 (basic reproduction number), β (transmission rate) and γ (average recovery rate) can be found in Batista [64] and Ranjan [65].

After carrying out the overall prediction, we particularly noted the predicted values of daily new cases from 6th May 2020 to 15th May 2020 in order to calculate rmse and do the required comparison with the hybrid model (Table 4).

To check the validity of the model on a longer duration, we trained the data for 200 days and predicted it for next 40 days. Since Arima is a linear model, we see that for the testing data, the graph just rises linearly up in Fig. 11.(a); similarly in 11.(b) we see the lstm model also settling down in the long run. But when the residual corrections are added in the hybrid arima, the graph shows some non linear variations in Fig. 11.(c). However, the non-linear variations also become constant over a period of time which goes to show that the error values captured in the training data more or start repeating over a period of time which is unlikely to happen in a real life scenario. Thus, to forecast for a longer duration, we may need to make proper adjustments in the model. Still compared to the single ARIMA and LSTM model, the hybrid model is more reliable (Table 5).

5. 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. 3). 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 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. First, the best model was selected for training the respective datasets on ARIMA 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. Both the models, ARIMA and NAR were then used to predict the future cases and residuals respectively. 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 (Eqs. (9)–(11)), we evaluated the prediction accuracy of both the models for all the three datasets. Reduced error as seen in Tables 3a–3c clearly advocate for the superiority of the proposed hybrid ARIMA model over a single ARIMA model. We have also compared the model with LSTM, SIR model and the hybrid ARIMA outperformes that as well.

Although our model has shown better performance compared to LSTM, SIR and ARIMA, the difference between the results however starts

to reduce as days increase for cumulative cases with larger dataset. This goes to show the limitation of our model to forecast on longer horizon of months. In addition to current covid transfer rate and prevention policies, uncertain behavioural patterns, and mitigation schemes also account for forecasting accuracy at longer intervals.

Still, our model is particularly suited for quick short term forecasts in an epidemic. 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.

6. 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. Our model also outperforms SIR and LSTM model for short term forecasts. Therefore, the new hybrid model can be considered as a reliable tool for policymakers in predicting short term forecasts of COVID-19 and devising proper strategies in due time.

However, for longer intervals, the difference of results between models reduces owing to the uncertainities of data, mitigation policies and behavioural patterns.

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

CRediT authorship contribution statement

Aman Swaraj: Methodology, Software. **Karan Verma:** Conceptualization. **Arshpreet Kaur:** Writing– original draft. **Ghanshyam Singh:** Visualization. **Ashok Kumar:** Investigation. **Leandro Melo Sales:** Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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.

References

[1] https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series, 2019.
[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] I.K. Oboho, et al., 2014 MERS-CoV outbreak in Jeddah—a link to health care facilities, N. Engl. J. Med. 372 (9) (2015) 846–854.
[6] World Health Organization, 2020. Coronavirus disease 2019 (COVID-19): situation report, 51.
[7] P. Zhou, X.L. Yang, X.G. Wang, et al., A pneumonia outbreak associated with a new coronavirus of probable bat origin, Nature (2020) [Epub ahead of print].
[8] Q. Li, X. Guan, P. Wu, 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, 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, Amol P. Bhondekar, Application of time series methods for dengue cases in North India (Chandigarh), J. Public Health (2019): 1-9.
[12] A. Forna, P. Nouvellet, I. Dorigatti, C. Donnelly, Case fatality ratio estimates for the 2013–2016 west African Ebola epidemic: application of boosted regression trees for imputation, Int. J. Infect. Dis. 79 (2019) 128.
[13] Kumar Shashvat, et al., Comparison of time series models predicting trends in typhoid cases in northern India, Southeast Asian J. Trop. Med. Public Health 50 (2) (2019) 347–356.
[14] S.-L. Jhuo, M.-T. Hsieh, T.-C. Weng, M.-J. Chen, C.-M. Yang, 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, K. VanderWaal, Identifying outbreaks of porcine epidemic diarrhoea virus through animal movements and spatial neighbourhoods, Sci. Rep. 9 (1) (2019) 1–12.
[16] G. Kalipe, V. Gautham, R.K. Behera, 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, 2018, December.
[17] R. Singh, R. Singh, A. Bhatia, Sentiment analysis using Machine Learning technique to predict outbreaks and epidemics, Int. J. Adv. Sci. Res 3 (2) (2018) 19–24.
[18] S.A. Abdulkareem, E.-W. Augustijn, T. Filatova, K. Musial, Y.T. Mustafa, “Risk perception and behavioural change during epidemics: Comparing models of individual and collective learning,”PloS one, 2020.
[19] T. Kuniya, 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, 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, K. Leung, G.M. Leung, Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study, Lancet 395 (2020) 689–697, 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, Int. J. Infect. Dis. (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] H.H. Elmousalami, A.E. Hassanien, Day Level Forecasting for Coronavirus Disease (COVID-19) Spread: Analysis, Modelling and Recommendations. ArXiv preprint arXiv:2003.07778, 2020.
[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, 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, Sonali Agarwal, COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms, medRxiv, 2020.
[33] D. Benvenuto, M. Giovanetti, L. Vassallo, S. Angeletti, M. Ciccozzi, 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] T. Dehesh, H.A. Mardani-Fard, P. Dehesh, Forecasting of COVID-19 Confirmed Cases in Different Countries with ARIMA Models, MedRxiv (2020).
[37] Gupta, Rajan, 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] Z. Shi, Y. Fang, Temporal relationship between outbound traffic from Wuhan and the 2019 coronavirus disease (COVID-19) incidence in China, MedRxiv (2020).
[41] Mohammed A.A. Al-Qaness, et al., Optimization method for forecasting confirmed cases of COVID-19 in China, J. Clin. Med. 9 (3) (2020) 674.
[42] Zeynep Ceylan, Estimation of COVID-19 prevalence in Italy, Spain, and France, Sci. Total Environ. 138817 (2020).
[43] Leila Moftakhar, S.E.I.F. Mozhgan, Marziyeh Sadat Safe, Exponentially Increasing Trend of Infected Patients with COVID-19 in Iran: A Comparison of Neural Network and ARIMA Forecasting Models, Iranian J. Public Health 49 (2020) 92–100.
[44] Y. Zhang, H. Yang, H. Cui, Q. Chen, Comparison of the Ability of ARIMA, WNN and SVM Models for Drought Forecasting in the Sanjiang Plain, China. Nat. Resour. Res. 29 (2019) 1447.
[45] X. Zhang, et al., Applications and comparisons of four time series models in epidemiological surveillance data, PLoS ONE 9 (2014), e88075.
[46] Q. Li, et al., Application of an autoregressive integrated moving average model for predicting the incidence of haemorrhagic fever with renal syndrome, Am. J. Trop. Med. Hyg. 87 (2012) 364–370.
[47] A.A. Adly, et al., Utilizing neural networks in magnetic media modelling and field computation: a review, J. Adv. Res. 5 (2014) 615–627.
[48] S. Haykin, Neural networks: a comprehensive foundation, 2nd ed., Prentice Hall, 1998.
[49] L. Ljung, System identification: theory for the user, 2nd ed., Prentice Hall PTR, 1998.
[50] J.T. Connor, R.D. Martin, L.E. Atlas, Recurrent neural networks and robust time series prediction, IEEE Trans. Neural Networks 5 (2) (1994) 240–254.
[51] Tugba Taskaya-Temizel, Matthew C. Casey, A comparative study of autoregressive neural network hybrids, Neural Networks 18 (5-6) (2005) 781–789.
[52] M.D. Philemon, Z. Ismail, J. Dare, A review of epidemic forecasting using artificial neural networks, Int. J. Epidemiologic Res. 6 (3) (2019) 132–143.
[53] Atilla Aslanargun, Mammadagha Mammadov, Berna Yazici, Senay Yolacan, Comparison of ARIMA, neural networks and hybrid models in time series: tourist arrival forecasting, J. Stat. Comput. Simul. 77 (1) (2007) 29–53.
[54] Ashu Jain, Avadhnam Madhav Kumar, Hybrid neural network models for hydrologic time series forecasting, Appl. Soft Comput. 7 (2) (2007) 585–592.
[55] L. Yu, 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 9 (2014), e98241.
[56] G.Peter Zhang, Time series forecasting using a hybrid ARIMA and neural network model, Neurocomputing 50 (2003) 159–175.
[57] K.W. Wang, et al., Hybrid methodology for tuberculosis incidence time-series forecasting based on ARIMA and a NAR neural network, Epidemiol. Infect. 145 (6) (2017) 1118–1129.
[58] Khalil Benmouiza, Ali Cheknane, Small-scale solar radiation forecasting using ARMA and nonlinear autoregressive neural network models, Theor. Appl. Climatol. 124 (3-4) (2016) 945–958.
[59] G.E.P. Box, G. Jenkins, Time Series Analysis, Forecasting and Control, Holden-Day, San Francisco, CA, 1970.
[60] C.W.J. Granger, Combining forecasts—Twenty years later, J. Forecasting 8 (1989) 167–173.
[61] K. Levenberg, A method for the solution of certain problems in least squares, Q. Appl. Math. 5 (1944) 164–168.
[62] D.J.C. MacKay, Bayesian interpolation, Neural Comput. 1992;4(3):415–47.
[63] Martin Fodslette Møller, A scaled conjugate gradient algorithm for fast supervised learning, Neural Networks 6 (4) (1993) 525–533.
[64] Milan Batista, Estimation of the final size of the covid-19 epidemic, medRxiv, doi, 10(2020.02):16–20023606, 2020.
[65] Ranjan, Rajesh, Predictions for COVID-19 outbreak in India using epidemiological models, MedRxiv, 2020.