SN Computer Science (2021) 2:372
https://doi.org/10.1007/s42979-021-00764-9

ORIGINAL RESEARCH

Performance Evaluation of Soft Computing Approaches for Forecasting COVID-19 Pandemic Cases

Muhammad Shoaib[1] · Hamza Salahudin[2] · Muhammad Hammad[3] · Shakil Ahmad[4] · Alamgir Akhtar Khan[5] · Mudasser Muneer Khan[6] · Muhammad Azhar Inam Baig[7] · Fiaz Ahmad[8] · Muhammad Kaleem Ullah[9]

Received: 24 February 2021 / Accepted: 2 July 2021 / Published online: 9 July 2021

Abstract

An unexpected outbreak of deadly Covid-19 in later part of 2019 not only endangered the economies of the world but also posed threats to the cultural, social and psychological barriers of mankind. As soon as the virus emerged, scientists and researchers from all over the world started investigating the dynamics of this disease. Despite extensive investments in research, no cure has been officially found to date. This uncertain situation rises severe threats to the survival of mankind. An ultimate need of the time is to investigate the course of disease transfer and suggest a future projection of the disease transfer to be enabled to effectively tackle the always evolving situations ahead. In the present study daily new cases of COVID-19 was predicted using different forecasting techniques; Autoregressive Integrated Moving Average (ARIMA), Exponential Smoothing/Error Trend Seasonality (ETS), Artificial Neural Network Models (ANN), Gene Expression Programming (GEP), and Long Short-Term Memory (LSTM) in four countries; Pakistan, USA, India and Brazil. The dataset of new daily confirmed cases of COVID-19 from the date on which first case was registered in the respective country to 30 November 2020 is analyzed through these five forecasting models to forecast the new daily cases up to 31st January 2020. The forecasting efficiency of each model was evaluated using well known statistical parameters R^2, RMSE, and NSE. A comparative analysis of all above-mentioned models was performed. Finally, the study concluded that Long Short-Term Memory (LSTM) neural network-based forecasting model projected the future cases of COVID-19 pandemic best in all the selected four stations. The accuracy of the model ranges from coefficient of determination value of 0.85 in Brazil to 0.96 in Pakistan. NSE value for the model in India is 0. 99, 0.98 in USA and Pakistan and 0.97 in Brazil. This high-accuracy forecast of COVID-19 cases enables the projection of possible peaks in near future in the aforementioned countries and, therefore, prove to be helpful in formulating strategies to get prepared for the potential hard times ahead.

Keywords Time-series forecasting · COVID-19 · Long Short-Term Memory (LSTM) · Autoregressive integrated moving average (ARIMA) · Exponential smoothing (ETS) · Artificial neural network (ANN) · Gene expression programming (GEP)

Introduction

An extremely infectious viral disorder, COVID-19, triggered by a novel coronavirus named SARS-CoV-2, arose in Wuhan, China in December 2019. After its speedy transmission form one country to another, on 12 March 2020, World Health Organization (WHO) announced COVID-19 a global pandemic [33]. The transfer of SARS-CoV-2 virus in humans is mainly driven by the tiny globules from human respiration, i.e., from talking, coughing and sneezing as well as from polluted surfaces [31]. The faster propagation of the virus is due to its ability to survive on different surfaces as long as 9 h at room temperature [5]. However, some recent research studies have concentrated on certain potential zoonoses mechanisms, such as other animals that may also possibly transmit COVID-19 [28]. This virus can induce severe respiratory disorder condition or sometimes multiple organs failure, which may escalate to medical collapse and even death of the infected person [10].

With this degree of severity and quick transmission, the number of COVID-19 confirmed cases had surpassed a massive mark of 50 million as of 30 November, 2020 [32],

✉ Muhammad Shoaib
msho127@aucklanduni.ac.nz

Extended author information available on the last page of the article

while the deaths were also recorded to be more than 1.5 million till the date. Despite the planning and execution of different measures including social distancing, lockdowns and precautions at national and internationals to mitigate the adverse impacts of the pandemic, there is still a continuously evolving situation. Researchers in the medical field from all over the world have been trying to figure out the cure but there is not recognized and registered effective drug against the virus. Simultaneously, many researchers and data scientists have been trying to accurately forecast the COVID-19 metrics using different data engineering and artificial intelligence approaches. An accurate forecast of the pandemic behavior and trend would help in effective planning and formulation of pandemic handling strategies to minimize the already aggravated economic, social and psychological impacts on at domestic, national and international levels in future.

Related Work

Several forecasting approaches have been implemented to study the future dynamics of COVID-19 pandemic in different parts of the world. These approaches include mathematical models, artificial intelligence approaches like Long Short-Term Memory (LSTM) models, autoregressive integrated moving average (ARIMA) technique, support vector regression (SVR), trust region reflective (TRR) algorithm and so on. Sarkar et al. [27] developed a mathematical model to forecasts the developments of COVID-19 situation in India. The model studies six parameters namely susceptible, asymptomatic, recovered, infected, isolated infected and quarantine susceptible, articulated as SARII q S q. Sensitivity analysis is performed to assess the effectiveness of model projections for parameter values and the sensitive parameters are calculated from the actual data on the COVID-19 pandemic in India. Their results demonstrate that the increasing infection rate can be significantly controlled by restricting the rate of interaction between infected and uninfected by quarantining the susceptible individuals. Moreover, it was also asserted that the combination of contact tracking and social distancing can be effective in controlling the ongoing pandemic. However, the study does not present any future projections of the pandemic course.

Likewise, another study by Pai et al. [20] implemented a mathematical approach based on susceptible–exposed–infectious–recovered (SEIR) model for forecasting the confirmed active cases of COVID-19 in India. The study also demonstrates the influence of national level lockdown in the country on active cases and aftermaths of lockdown removal. They predicted an inflation of up to 21 percent in the peak active cases in response to different hypothetical situations of relaxation or normalization in control strategies. Moreover, the authors suggested another 40-day national level lockdown to flatten the increasing graph of active cases in India. Nabi [18] also implemented a susceptible–exposed–symptomatic infectious–asymptomatic infectious–quarantined–hospitalized–recovered–dead (SEI$_D$I$_U$QHRD) compartmental model based on trust region reflective (TRR) algorithm to study the dynamics of the pandemic. They predicted the daily confirmed active cases peaks in Bangladesh, India, Brazil and Russia. Moreover, authors also suggested that relaxation in lockdown or social distancing measures can quickly intensify the pandemic outbreak. Farooq and Bazaz [7] employed deep learning technique to propose an artificial neural network- (ANN) based simulation model. They implemented population compartmentalizing approach to divide the population into two subsets: high-risk (HR) and low-risk (LR) compartments. After subjection of the pandemic dynamics to the population subsets, it was suggested that if HR subset practices self-isolation and allows the LR subset to gain immunity. Then on release, HR subset would be safe from infectious surroundings rather it will be surrounded by already immunized LR subset. Thus, reducing the risk of further escalation of active cases. Ribeiro et al. [25] used ARIMA, SVR, random forest (RF), cubist regression (CUBIST), ridge regression (RIDGE), stack-assembling for time series forecasting of cumulative cases in Brazil. They made forecasts with one, three, and six days ahead with forecasting errors in the range of 0.87–3.51 percent, 1.02–5.63 percent, and 0.95–6.90 percent, respectively. After comparative analysis of model performance, it was concluded that SVR model out-performed all other models used in the study. Hybrid machine learning approaches of adaptive neuro-fuzzy inference system (ANFIS) and multi-layered perceptron–imperialist competitive algorithm (MLP-ICA) were used by Pinter et al., [23] to predict COVID-19 outbreak in Hungry. This study recommends machine learning may be considered as an alternative of standard epidemiological models, i.e., susceptible–infected–resistant (SIR)-based models to model the pandemic outbreak. Singh et al. [29] used advanced autoregressive moving average model to find the top 15 countries with spatial mapping of the COVID-19-confirmed cases. The developed model was also used for predicting the COVID-19 disease spread trajectories for the next 2 months. A novel algorithm which make use of machine learning (ML) and evolutionary computation (EC) was proposed by Khalilpourazari and Hashemi [13] to model and predict the COVID-19 pandemic in Quebec, Canada. Roy et al. [26] investigated using machine learning techniques to characterize the effect of COVID-19 pandemic worldwide. An additive regression model with interpretable parameters was proposed in the study. The study performed an accurate analysis of country-wise as well as province/state-wise confirmed cases, recovered cases, deaths, prediction of pandemic viral attack and how far it is expanding globally. Different machine learning models were employed by Malki et al. [16] for predicting the spread of coronavirus using the weather data. The machine learning models used in the study includes linear models (Linear Regression, Lasso Regression, Ridge Regression, Elastic Net, Least Angle Regression, Lasso Least Angle Regression, Orthogonal Matching Pursuit, Bayesian Ridge, Automatic Relevance Determination, Passive Aggressive Regressor, Random Sample Consensus, TheilSen Regressor, Huber Regressor), ensemble models (Random Forest, Extra Trees Regressor, AdaBoost Regressor, Gradient Boosting Regressor) Extreme Gradient Boosting, Light Gradient Boosting, Support Vector Machine (SVM), K-Nearest Neighbors Regressor, Multilevel Perceptron (MLP) and Decision Tree.

Similarly, many other studies have extracted, studied, modeled or forecasted different features of pandemic course using different methodologies [1, 3, 5, 6, 12, 15, 17, 19, 21, 22, 24, 28].

Materials and Methods

This section contains the brief description of forecasting techniques used in the study and the dataset of daily confirmed cases of COVID-19 in Pakistan, USA, Brazil and India.

Dataset

There were four different countries named Pakistan, USA, Brazil and India in the present study as they have high number of COVID-19 cases. The dataset of new daily confirmed cases of COVID-19 from the date on which first case was registered in the respective country to 30 November 2020 were extracted from https://ec.europa.eu/eurostat. The database can be accessed freely and extracted easily.

Long Short-Term Memory (LSTM)

Hochreiter and Schmidhuber [11] first developed LSTM, a deep learning artificial recurrent neural network model, to solve gradient problem drawbacks related to simple recurrent neural networks (RNN). This deep learning model is invaluable to generate certain significant insight into complicated problems such as forecasting with time series, speech detection and text recognition. The conventional RNN model is not able to recall the effect of the initial values in the data after a specific sequence duration of just ten to fifteen steps [14]. This implies that historical rainfall for just 10–12 days will have a real effect on the forecast rainfall. This effect will be weakened as the series expands and at a certain limit forecast will be produced with previous rainfall forecast that decrease the precision of the model. In contrast to RNN, LSTM implies that supplemental information and data processing gates are used in each memory cells. The memory cells are stored in special units named as memory blocks in hidden layers. This results in the effective back propagation of the gradient by identity function; therefore, the gradient getting backpropagated does not burst or disappear but stays stable over the span of the sequence and thus, the effect of the early phases remains unchanged. Even so, the application of LSTM in time series forecasting is very restricted due to the complicated nature of the model and the requirement for more computing duration and high-end equipment / software resources. The entry into a single memory cell is the cell state of the former cell (*Ct*−1), the hidden state of the former cell (*ht*-1) as well as the input vector (*Xt*) (*Xt*). The *Xt* and *ht*−1 inputs are fed into filter gates. Sigmoid non-linearity layer (σ) examines *Xt* and *ht*−1 and condenses the input elements of *Ct*−1 to a limit of [0,1] defining how much of each actual valued input will be transferred along. Zero value means "pass nothing through the gate" and one implies "leave nothing through the gate" anything between null and one implies that the percentage of the input variable will be allowed to pass. Tanh non-linearity generates a unique candidate vector from *Xt* and h_t−1 inputs by condensing the input elements to the range [− 1,1]. This candidate vector is transferred through the input gate (*it*) and then connected to the former cell state (*Ct*−1) which has already progressed through the forget gate (*ft*). As a result, a new cell state (*Ct*) is created for timestep *t*.

This modified cell state is transferred through the output gate after tanh non-linearity is implied. It produces the final output of the memory, also called the hidden state (*ht*). A set of mathematical equations that describe the method at each stage is given below [source: Hochreiter and Schmidhuber [11]]:

$$\sigma = \frac{e^x}{(1 + e^x)},$$

$$\tan h = \frac{e^x - e^{-x}}{e^x + e^{-x}},$$

$$i_t = \sigma\left(W_i\left[X_t, h_{t-1}\right] + \theta_i\right),$$

$$f_t = \sigma\left(W_f\left[X_t, h_{t-1}\right] + \theta_f\right),$$

$$o_t = \sigma\left(W_o\left[X_t, h_{t-1}\right] + \theta_o\right),$$

$$\acute{C}_t = \tan h\left(W_c\left[X_t, h_{t-1}\right] + \theta_c\right),$$

$$C_t = f_t * C_{t-1} + i_t * \acute{C}_t.$$

Figure 1 shows the structure of a simple LSTM model.

Artificial Neural Network Models (ANN)

Artificial Neural Network (ANN) is a bio-inspired computational technique for modeling a wide variety of non-linear structures. The high degree of precision obtained by ANN is due to the concurrent processing of information across the neurons network and the connected weights. The network structure mainly comprises of three consented layers. First layer is input layer which contains input neurons that are connected with hidden neurons of hidden layers through connecting weights. Neurons in hidden layer are further connected with output layer neurons. The structure of ANN models enables flow of information in forward direction from input layer toward output layer and backflow of modeling error from output layer to input layer. The simplest structure of ANN model is referred as to Feedforward Neural Network (FFNN) due to its ability to move information in forward direction as mentioned above. A general ANN model can be mathematically described as:

$$x_{\text{out}} = f_2\left[\sum_{j=1}^{h} W_{kj} f_1\left(\sum_{i=1}^{n} W_{ji} x_i + \theta_{jo}\right) + \theta_{ko}\right],$$

where, x_i is input value to ith input layer neuron, y_{out} is the output at kth output layer neuron, f_1 is non-linear activation function for hidden layer and f_2 is linear activation function for output layer. n and h represent the number of neurons in input layer and hidden layer, respectively. θ_{jo} and θ_{ko} are the bias units for jth input layer neuron and kth output layer neuron, while W_{kj} is the weight connecting jth hidden layer neuron and kth output layer neuron, and W_{ji} is the weight connecting input layer neuron i and hidden layer neuron j.

Temporal correlation plays a very important role in historic time series. Including time component in neural network directly or indirectly can improve its performance and enhance accuracy [9]. So, the inputs to ANN are lagged to an order of 1–10 and model performance is evaluated for each lag order. Lag order with best performance is considered as optimum lag while optimum lag order is cross-validated with error autocorrelation graphs and significant lag number is considered with optimum lag order.

Autoregressive Integrated Moving Average (ARIMA)

ARIMA is the most common and widely used model for time series forecasting, the main objective of this model is to forecast future values by using the past values. ARIMA is also called the Box–Jenkins process. Because of its generality, it is prominent, and it can be used with or without seasonal elements. ARIMA consist of two major processes autoregressive process (AR) and moving average process (MA). A typical ARIMA model can be written as ARIMA (p, d, q), where p = order of auto-regression (AR), d = order of integration (also known as differencing) and q = order of

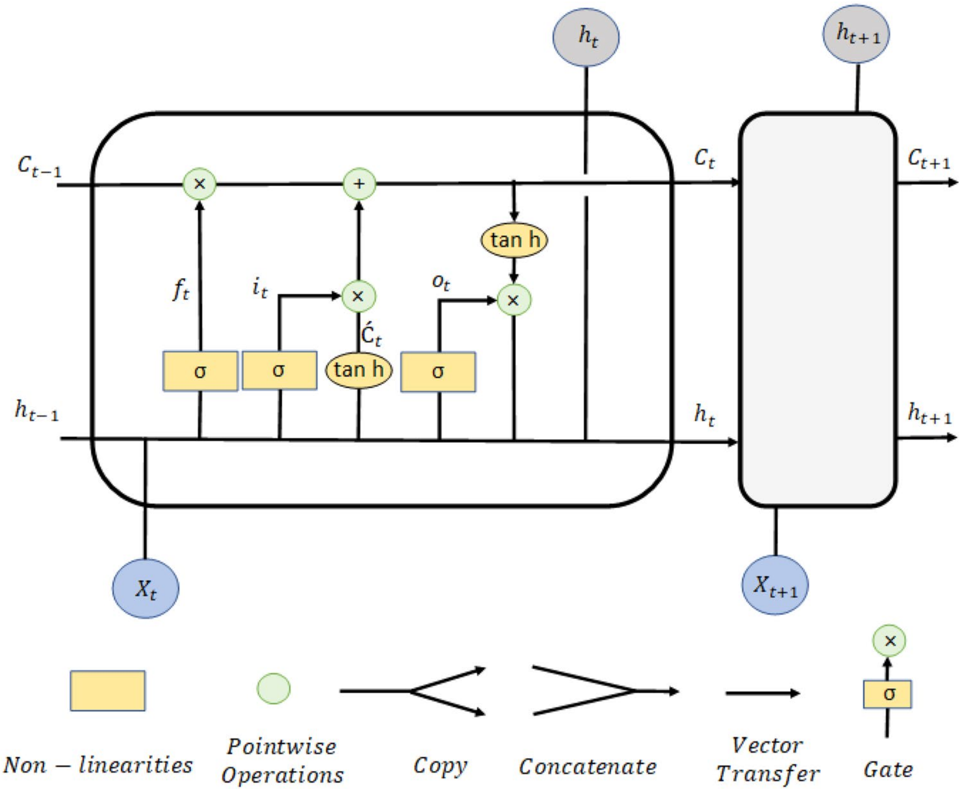


Fig. 1 Structure of LSTM model [source: Duong and Bui [30]]

moving average (MA). In certain cases, ARIMA models are used if data indicates non-stationarity in the mean context. In model construction, ARIMA has four main phases which are identification, assessment, diagnostic and prediction. To construct an ARIMA model, stationarity is an essential condition that would be useful in prediction. Data processing is done to make the time series stationary in the identification process. In addition, values of p and q are determined in the identification step by using unit root test, ACF and PACF. In assessment step the suitable ARIMA model is estimated using p, d and q values. In diagnostic phase the residuals are checked to look white noise for choosing the best ARIMA model having well-behaved residuals. Finally, forecasting is obtained in the form of set aside last few data points. These phases are presented in the Fig. 2.

Exponential Smoothing/Error Trend Seasonality (ETS)

Exponential smoothing (ETS) is a prediction tool for single variable in time series forecasting. It is an efficient prediction method that can be used as a substitute to the most common technique Box–Jenkins. ETS is a powerful format for constructing a smooth time series. Exponential smoothing gives declining weights exponentially as the spectrum grows older, although the previous observations are equally weighted in moving averages. There are three types of exponential smoothing: single, double and triple. In this study, tripe exponential smoothing is utilized to forecast daily new COVID-19 cases because it is suitable for seasonally or other recurrent non-linear data models. The equation of a simple ETS can be written as:

$$y_o = x_o,$$

$$y_t = \alpha x_{t-1} + (1-\alpha) y_{t-1}.$$

Here, y_t is the output of ETS, x_t represents the raw data at the beginning and α is the smoothing factor, the value of which varies from 0 to 1.

Gene Expression Programming (GEP)

Gene expression programming is a transformed form of genetic programming (GP) and genetic algorithm (GA) [8]. GP is the general form derived from the genetic algorithm. Jone Koza first constructed a computer-based model for GP in 1988 to overcome the issue using the Darwinian selection principle [2]. GP is an artificial intelligence-based predictive technique that generates a framework that replicates the development of living organisms. The GEP model has 5 parameters, which are: fitness function, set of terminals, parameters of control, terminal conditions, and set of functions [4]. The GEP method generates a population set of randomly chosen individual genes and then transforms each entity into an expression tree of various types to represent their numerical form solutions. Then, the target is correlated with the estimated one, and each particular entity's fitness score is calculated. The system stops if the model provides a better performance. The best longevity genes from individuals are obtained and transferred on to the next generation. This cycle continues until it achieves the optimal survival gene with an adamant fitness score. Following are the simple calculation procedures for the GEP Fig. 3.

1. For accurate classification arrange the fitness function, the fitness function of any GEP entity i can be expressed as $\sum_{k=1}^{N}\left(P_{ik}=T_k\right)=\text{Fitness}_i$, where, N represents the total number of COVID-19 cases, T_k represents the target value under GEP individual i, and P_{ik} refers to the cases prediction under GEP individual i.
2. This phase randomly produces a fixed length chromosome for each individual for the initial population
3. Chromosomes are then express in the tree expression form and determine each individual's fitness.
4. Based on their fitness value, perform replication and revision and choose the most efficient entity.

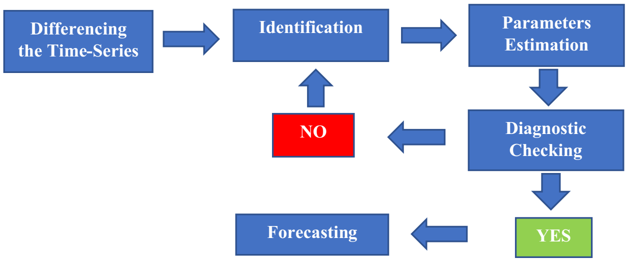
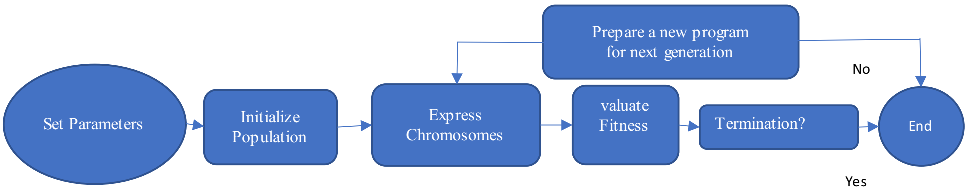


Fig. 2 Block diagram of ARIMA model building process

Fig. 3 Block diagram of GEP model building process



5. Repeat the above phases (2–4) on the basis of a specified number of generations.

Performance Parameters

To test the performance of established models (ARIMA, ETS and LSTM), many statistical parameters can be used. In the present study the effectiveness of each model is calculated with three statistical measures, namely Nash–Sutcliffe efficiency (NSE), determination coefficient (R^2) and root mean squared error (RMSE). Nash–Sutcliffe efficiency (NSE) introduced by Sutcliffe and Nash is one of the principle most used in model performance assessment.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(C_A - C_P\right)^2},$$

$$\mathrm{NSE} = \left[1 - \frac{\sum_{i=1}^{n}\left(C_A - C_P\right)^2}{\sum_{i=1}^{n}\left(C_A - \overline{C_A}\right)^2}\right], \tag{2}$$

$$R^2 = \left[\frac{\sum_{i=1}^{n}\left(C_A - \overline{C_A}\right)\left(C_P - \overline{C_P}\right)}{\sqrt{\sum_{i=1}^{n}\left(C_A - \overline{C_A}\right)^2 \sum_{i=1}^{n}\left(C_P - \overline{C_P}\right)^2}}\right]^2,$$

where C_A and C_P denote the actual daily new cases and predicted daily new cases of COVID-19. $\overline{C_A}$ and $\overline{C_P}$ represent the corresponding mean of daily new cases values. Higher values are more desirable and ideal value is nearer to 1 in both R^2 and NSE, while in case of RMSE with the exception the desirable value is nearer to 0. The scale of R^2 and NSE is from 0 to 1. Zero value indicate no relation between actual and observed cases, while 1 describes even and continuous linear relationship.

Results and Discussion

The dataset of new daily confirmed cases of COVID-19 from the date on which first case was registered in the respective country to 30 November 2020 is analyzed through five different forecasting models to forecast the new daily cases up to 31st January 2020. To check the accuracy of these models, the dataset is divided into two parts: training and testing. First of all, observation till 30 September 2020 is used as the training data to forecast the new daily cases from 1st October 2020 to 30 November 2020. Actual cases and the predicted cases of this duration are then compared to evaluate the precision of the forecasting models using the above-mentioned statistical parameters (NSE, R^2, and RMSE). Complete summary for testing phase is shown in Table 1.

In Fig. 4, the performance of ARIMA, ETS, LSTM, ANN and GEP forecasting models is evaluated in terms of root mean squared error (RMSE), Nash Sutcliffe Efficiency (NSE), and coefficient of determination (R^2). The x-axis in the graph represents forecasting techniques in all the four selected countries and the y-axis shows RMSE (no. of cases), R^2, and NSE (%). From Fig. 4, it can be clearly seen that LSTM- and ANN-based forecasting models give the best results as compared to ARIMA, ETS and GEP models in all the four selected countries.

LSTM and ANN exhibited the lowest values of RMSE, i.e., 177 and 128, 4231 and 2529 cases in Pakistan and Brazil, respectively. While, in USA and India, where cumulative cases are relatively more than other two countries, GEP outperforms LSTM models in terms of root mean squared error. GEP and ANN yield 10,990 and 9107, 3236 and 2529 in USA and India, respectively. Similarly, LSTM and ANN exhibited the highest values of the coefficient of determination and NSE as compared to ARIMA and ETS. It can be clearly observed that ARIMA, ETS and GEP (in some cases) resulted in higher RMSE and lower NSE which indicates an overall less accurate performance of these models. Figure 5 shows the forecasting of daily new COVID-19 cases in Pakistan using different forecasting techniques. It can be seen in Fig. 5a and b that there is a big gap between actual cases and the testing lines

Table 1 Summary of the testing phase

Country	Model	RMSE	R^2	NSE (%)
Pakistan	ARIMA	1246	0.77	49
	ETS	2163	0.56	53
	LSTM	177	0.96	98
	ANN	128	0.98	99
	GEP	186	0.96	98
USA	ARIMA	31,449	0.82	90
	ETS	46,946	0.64	79
	LSTM	14,564	0.95	98
	ANN	9107	0.96	99
	GEP	10,990	0.94	98
Brazil	ARIMA	21,772	0.23	34
	ETS	7131	0.59	92
	LSTM	4231	0.85	97
	ANN	6865	0.81	93
	GEP	7939	0.46	91
India	ARIMA	5695	0.81	98
	ETS	9462	0.82	96
	LSTM	3901	0.91	99
	ANN	2529	0.97	99
	GEP	3236	0.94	99

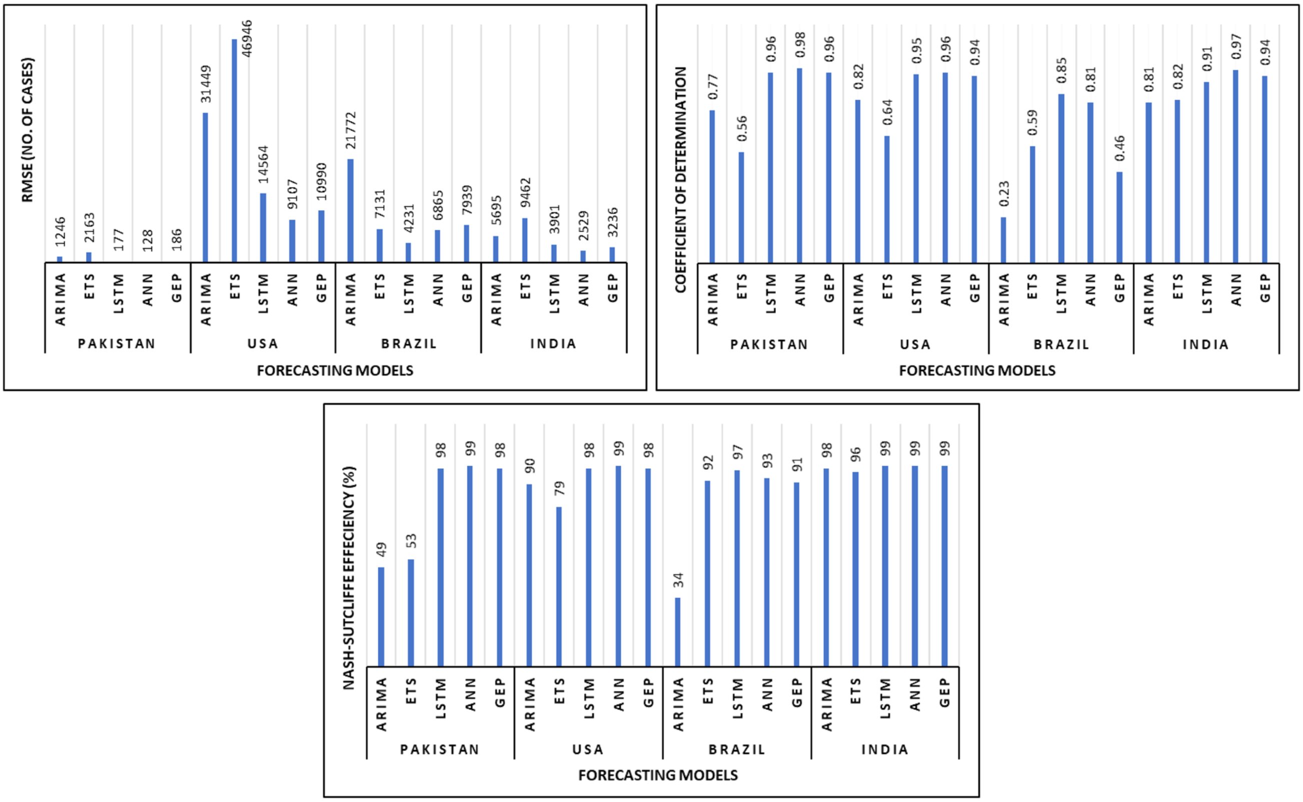


Fig. 4 Graphical representation of the summary of the testing phase

using ARIMA and ETS models, because of the poor testing results the forecasting through these techniques are not reliable. Although the results of LSTM are far better, LSTM models performed well in making daily new cases forecasts with the best accuracy rate at all four stations (Fig. 5c). It can also be observed from Fig. 5d and e that both ANN and GEP models performed better in testing phase as testing and actual cases lines closely coincide to each other for both models but forecasting from both model, ANN and GEP, does not seem realistic and reflects behavior same as ARIMA and ETS forecasts. The drop in forecasting performance of ANN and GEP models implies that LSTM models are best considered for COVID-19 cases future projections in Pakistan.

Figure 6 shows the forecasting of daily new cases in Brazil using different forecasting techniques. It can be clearly observed from Fig. 6a that ARIMA performed with the lowest accuracy rate among other methods. LSTM exhibited the highest values of coefficient of determination and NSE as compared to ARIMA and ETS as shown in Fig. 6b and Fig. 6c. The actual cases and testing lines are excellently close together using LSTM model which results in a trustworthy forecast. While GEP and ANN models produce similar results with good performance in testing but drastic drop in forecasting step (Fig. 6d, e). Consistency of LSTM models in training, testing and forecasting steps makes the technique superior.

Similarly, Figs. 7 and 8 show the forecasting of daily new cases in USA and India, respectively, using different forecasting techniques. It can be clearly observed that ARIMA and ETS resulted in higher RMSE and lower NSE as shown in Fig. 4 and Table 1. This indicates an overall less accurate performance of these models. The LSTM exhibited the highest values of coefficient of determination and NSE as compared to ARIMA and ETS as shown in Fig. 4 and Table 1. Graphically representation shows (Figs. 7a, b and 8a, b) that the actual cases and predicted cases are not together enough at both stations, while in the case of LSTM these lines are perfectly matching (Figs. 7c, 8c) as LSTM has 98% NSE value at all the selected four stations (Fig. 4). Nash–Sutcliffe efficiency for GEP and ANN is higher for both the countries, USA, and India (Fig. 4), but during forecasting both models produced poor forecasts with less realistic projection of future cases in USA and India (Figs. 7d, e and 8d, e). These forecasted results are valid till no vaccination is available for COVID-19 and all four countries do not significantly change their strategy against the pandemic. If any country changes its strategy drastically, the actual cases during the forecasting period may vary from the forecasted results. GEP models

Fig. 5 Forecasting of daily new cases in Pakistan using various soft computing approaches

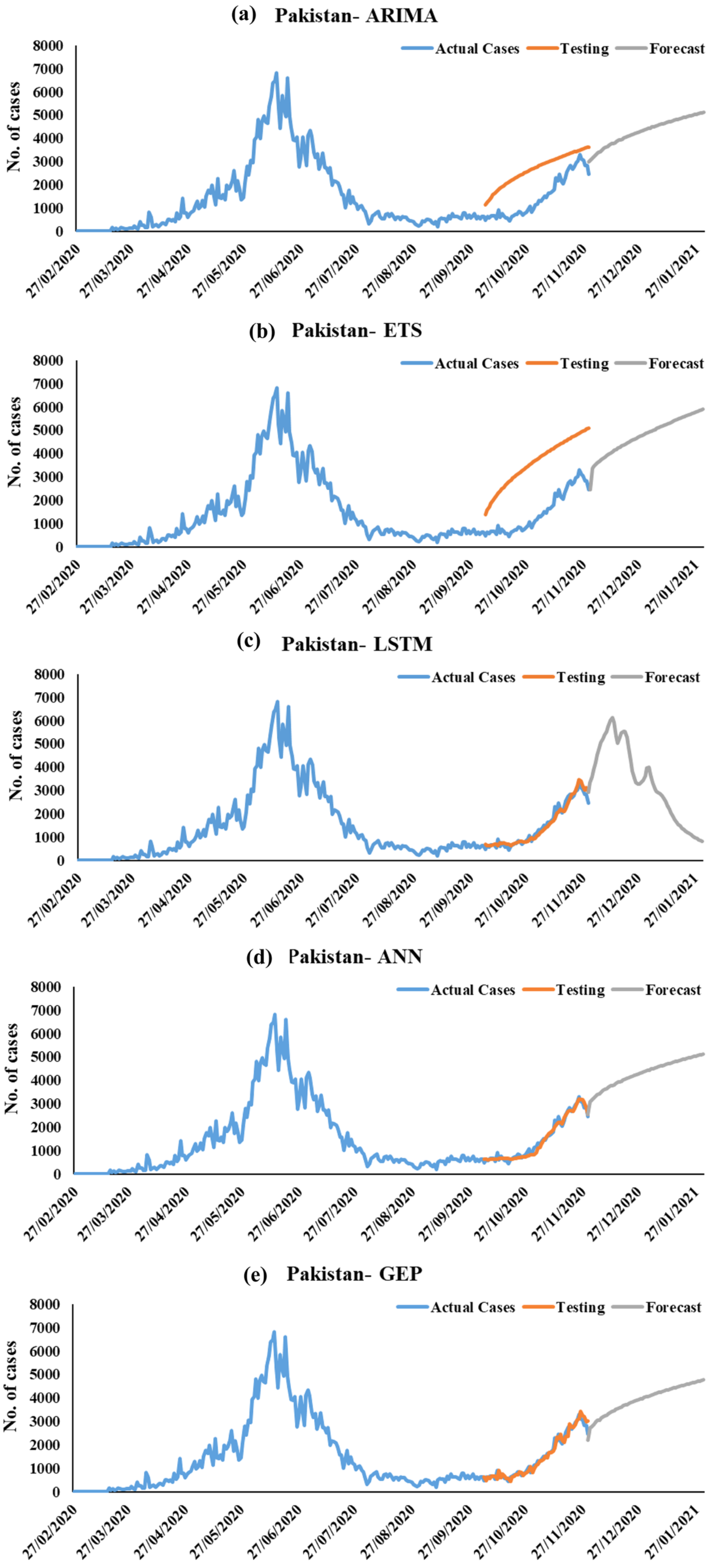
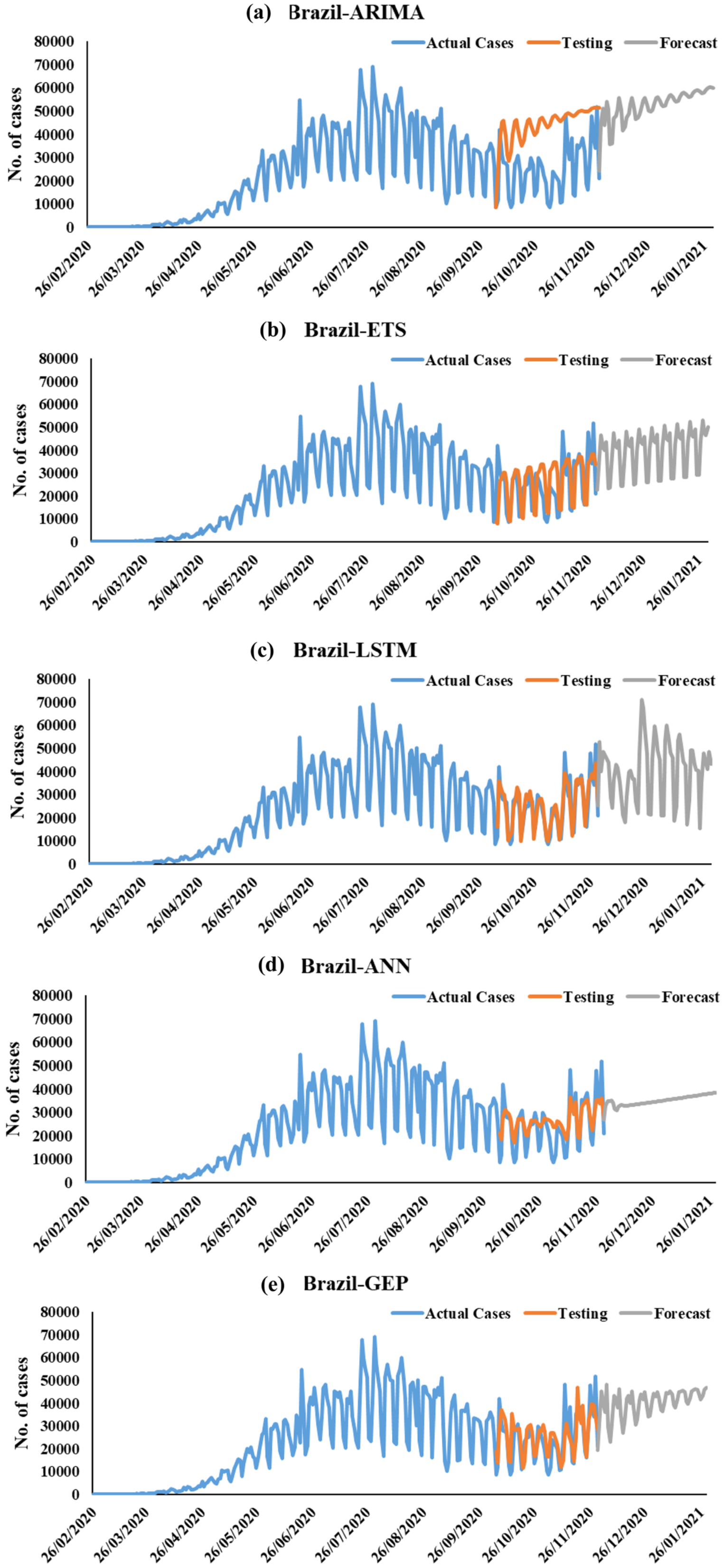


Fig. 6 Forecasting of daily new cases in Brazil using various soft computing approaches



(a) Brazil-ARIMA

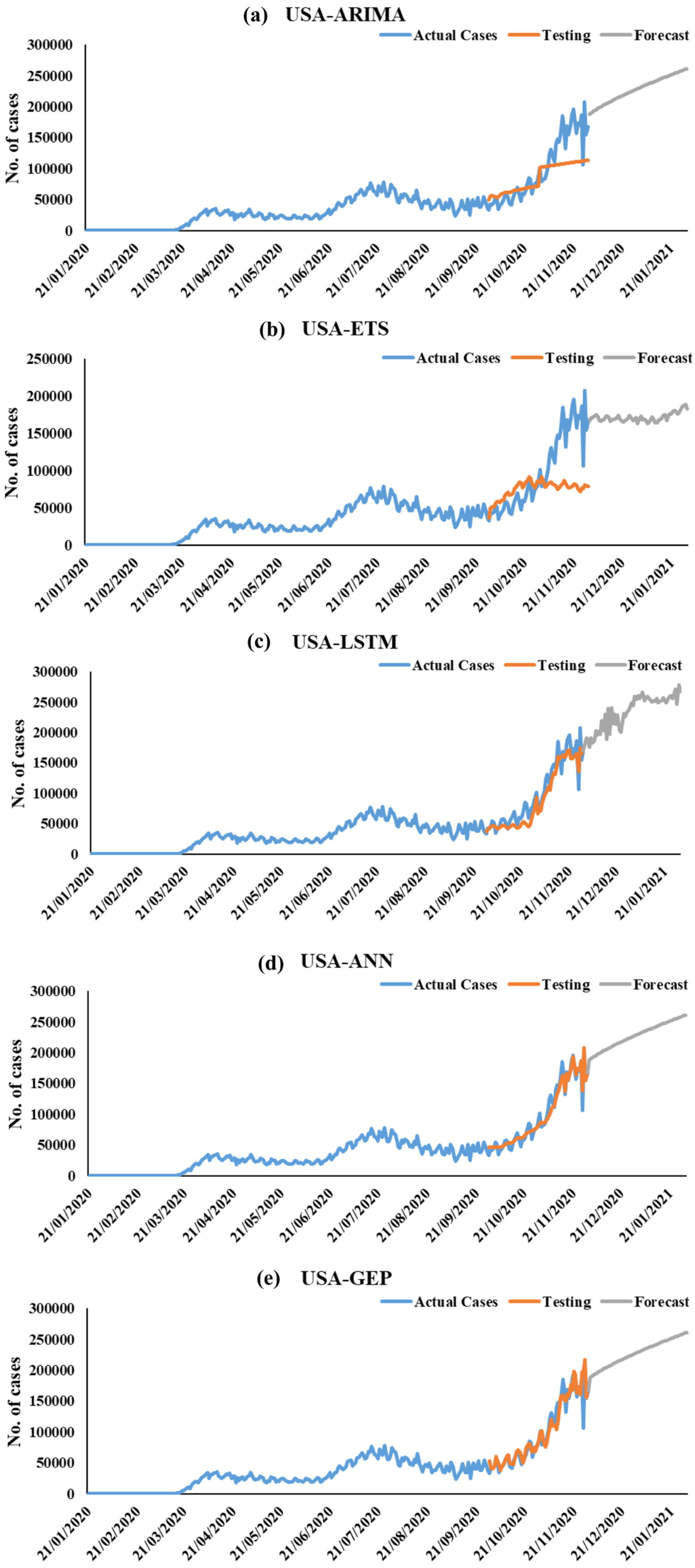
(b) Brazil-ETS

(c) Brazil-LSTM

(d) Brazil-ANN

(e) Brazil-GEP

Fig. 7 Forecasting of daily new cases in USA using various soft computing approaches



(a) USA-ARIMA

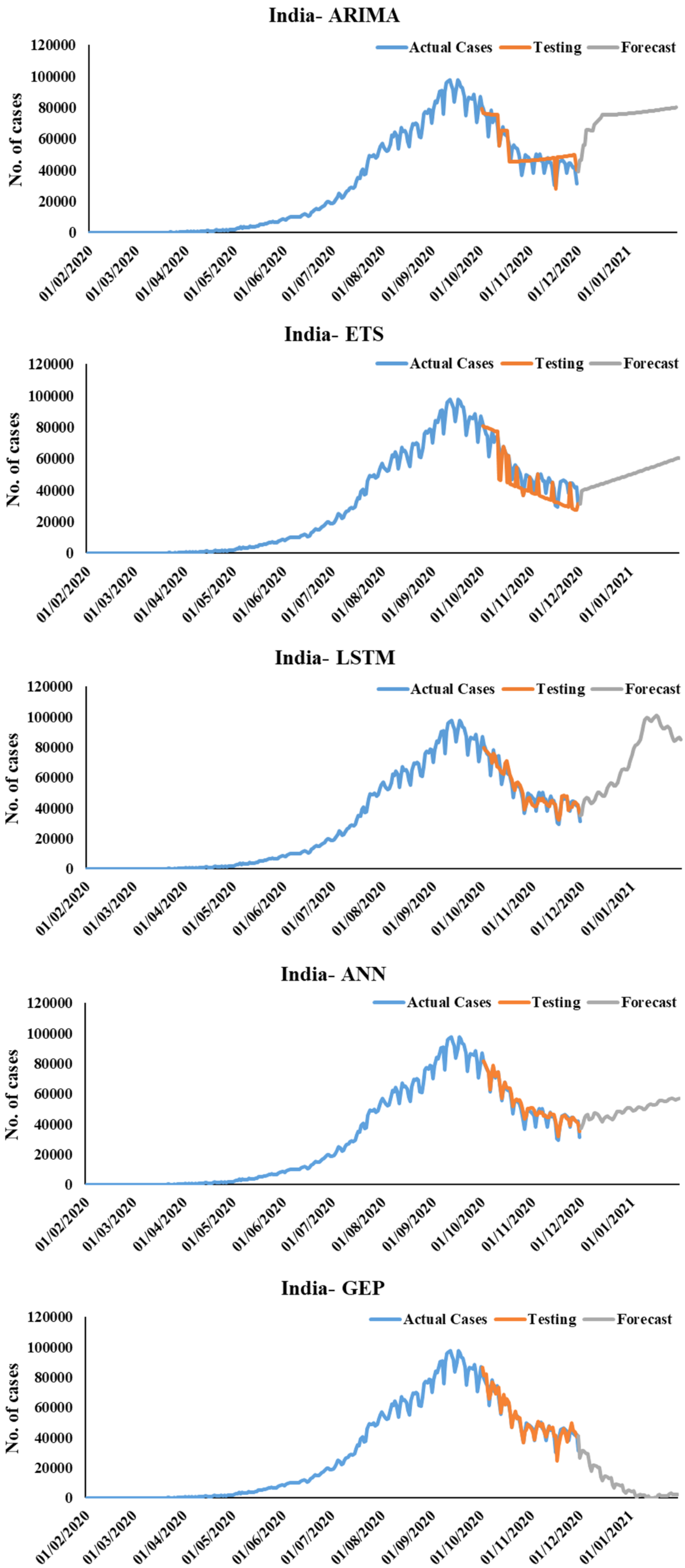
(b) USA-ETS

(c) USA-LSTM

(d) USA-ANN

(e) USA-GEP

Fig. 8 Forecasting of daily new cases in India using various soft computing approaches



India- ARIMA

India- ETS

India- LSTM

India- ANN

India- GEP

forecast COVID-19 cases to be decrease to almost zero by the end of forecasting period while LSTM forecast projects another peak during the later part of forecasting period which tends to decrease ahead.

Conclusion

In the whole world, COVID-19 infections are spreading rapidly and there is a need for robust preventive measures in the near future. For better control and prevention proper forecasting of new confirmed cases is vital. Therefore, in this study, an approach for the prediction of the daily new cases of COVID-19 pandemic was proposed across Pakistan, Brazil, India and the United States. The dataset of new daily confirmed cases from the date on which first case was registered in the respective country to 30 November 2020 was utilized to forecast the data points till 31st January 2020. A comparative analysis was performed between conventional forecasting techniques (Artificial Neural Network Models ANN, Gene Expression Programming GEP, Autoregressive Integrated Moving Average ARIMS and Exponential Smoothing ETS) and Long Short-Term Memory LSTM. To check the accuracy of these models, observation till 30 September 2020 is used as the training data to forecast the new daily cases from 1st October 2020 to 30 November 2020. Present cases and expected cases are then compared to validate the prediction models with statistical parameters (NSE, R^2, and RMSE). Finally, the results revealed that ARIMA and ETS resulted in higher RMSE and lower NSE values which indicates an overall less accurate performance of these models and LSTM exhibited the highest values of coefficient of determination and NSE (98%) as compared to ARIMA and ETS. NSE and Coefficient of determination values for ANN and GEP were also observed to be equal, better or competent to LSTM models but during forecasting only LSTM models produced realistic forecasts while the performance of ANN and GEP models significantly dropped. Therefore, LSTM can be successfully used to forecast daily new confirmed cases of COVID-19. These results may be very helpful if a vaccine is not readily available and if lockdown cannot be economically feasible in any region, for policymakers around the world to minimize the number of deaths.

Funding No funding information is available.

Availability of Data and Material Data are available on World Health Organization (WMO) website.

Code Availability MATLAB and DTREG are used for the study.

Declarations

Conflict of interest No conflict of interest.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication The authors hereby grant all rights of publication of the manuscript to the publisher.

References

1. Anastassopoulou C, Russo L, Tsakris A, Siettos C. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. PLoS ONE. 2020;15(3):1–21. https://doi.org/10.1371/journal.pone.0230405.
2. Aslam F, Farooq F, Amin MN, Khan K, Waheed A, Akbar A, Javed MF, Alyousef R, Alabdulijabbar H. Applications of gene expression programming for estimating compressive strength of high-strength concrete. Adv Civil Eng. 2020. https://doi.org/10.1155/2020/8850535.
3. Basu S, Campbell RH. Going by the numbers : Learning and modeling COVID-19 disease dynamics. Chaos Solitons Fractals. 2020;138: 110140. https://doi.org/10.1016/j.chaos.2020.110140.
4. Cheng CH, Chan CP, Yang JH. A seasonal time-series model based on gene expression programming for predicting financial distress. Comput Intell Neurosci. 2018;2018(1):1067350. https://doi.org/10.1155/2018/1067350.
5. Doremalen NV, Bushmaker T, Morris DH, Holbrook MG, Gamble A, Williamson BN, et al. Aerosol and Surface stability of SARS-CoV-2 as compared with SARS-CoV-1 | enhanced reader. N Engl J Med. 2020;382(16):1564–7. https://doi.org/10.1056/NEJMc2004973.
6. Elsheikh AH, Saba AI, Elaziz MA, Lu S, Shanmugan S, Muthuramalingam T, Kumar R, Mosleh AO, Essa FA, Shehabeldeen TA. Deep learning-based forecasting model for COVID-19 outbreak in Saudi Arabia. Process Saf Environ Prot. 2021;149:223–33. https://doi.org/10.1016/j.psep.2020.10.048.
7. Farooq J, Bazaz MA. A novel adaptive deep learning model of Covid-19 with focus on mortality reduction strategies. Chaos Solitons Fractals. 2020. https://doi.org/10.1016/j.chaos.2020.110148.
8. Ferreira C. Gene expression programming in problem solving. Soft Comput Ind. 2002;1996:635–53. https://doi.org/10.1007/978-1-4471-0123-9_54.
9. French MN, Krajewski WF, Cuykendall RR. Rainfall forecasting in space and time using a neural network. J Hydrol. 1992;137:1–31.
10. Gibson PG, Qin L, Puah S. COVID-19 ARDS: clinical features and differences to "usual"pre-COVID ARDS. Med J Aust. 2020;213(2):54–6.
11. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computat. 1997;9(8):1735–80.
12. Kapoor A, Ben X, Liu L, Perozzi B, Barnes M, Blais M, O'Banion. Examining COVID-19 forecasting using spatio-temporal graph neural networks. 2020.
13. Khalilpourazari S, Hashemi Doulabi H. Designing a hybrid reinforcement learning based algorithm with application in prediction of the COVID-19 pandemic in Quebec. Ann Oper Res. 2021. https://doi.org/10.1007/s10479-020-03871-7.
14. Kratzert F, Klotz D, Brenner C, Schulz K, Herrnegger M. Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrol Earth Syst Sci. 2018;22:6005–22.

15. Li Q, Feng W, Quan YH. Trend and forecasting of the COVID-19 outbreak in China. J Infect. 2020;80(4):469–96. https://doi.org/10.1016/j.jinf.2020.02.014.
16. Malki Z, Atlam E-S, Hassanien AE, Dagnewd G, Elhosseini MA, Gadb I. Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. Chaos Solitons Fractals. 2020;138: 110137.
17. Martelloni G, Martelloni G. Modelling the downhill of the Sars-Cov-2 in Italy and a universal forecast of the epidemic in the world. Chaos Solitons Fractal. 2020;139: 110064. https://doi.org/10.1016/j.chaos.2020.110064.
18. Nabi KN. Forecasting COVID-19 pandemic: adata-driven analysis. Chaos Solitons Fractals. 2020;139:15. https://doi.org/10.1016/j.chaos.2020.110046.
19. Niazkar HR, Niazkar M. Application of artificial neural networks to predict the COVID-19 outbreak. Global Health Res Policy. 2020. https://doi.org/10.1186/s41256-020-00175-y.
20. Pai C, Bhaskar A, Rawoot V. Investigating the dynamics of COVID-19 pandemic in India under lockdown. Chaos Solitons Fractals. 2020. https://doi.org/10.1016/j.chaos.2020.109988.
21. Perc M, GorišekMiksić N, Slavinec M, Stožer A. Forecasting COVID-19. Front Phys. 2020;8:1–5. https://doi.org/10.3389/fphy.2020.00127.
22. Petropoulos F, Makridakis S, Stylianou N. Forecasting COVID-19 confirmed cases and deaths with a simple time-series model. Int J Forecast. 2020. https://doi.org/10.1016/j.ijforecast.2020.11.010.
23. Pinter G, Felde I, Mosavi A, Ghamisi P, Gloaguen R. COVID-19 pandemic prediction for hungary; a hybrid machine learning approach. Mathematics. 2020;8(6): 890. https://doi.org/10.3390/math8060890.
24. Pinson P, Makridakis S. Pandemics and forecasting: the way forward through the Taleb-Ioannidis debate. Int J Forecast. 2020. https://doi.org/10.1016/j.ijforecast.2020.08.007.
25. Ribeiro MHDM, da Silva RG, Mariani VC, dos Coelho LS. Short-term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil. Chaos Solitons Fractals. 2020. https://doi.org/10.1016/j.chaos.2020.109853.
26. Roy A, Jose J, Sunil A, Gautam N, Nathalia D, Suresh A. Prediction and spread visualization of covid-19 pandemic using machine learning. Preprints. 2020. https://doi.org/10.20944/preprints202005.0147.v1.
27. Sarkar K, Khajanchi S, Nieto JJ. Modeling and forecasting the COVID-19 pandemic in India. Chaos Solitons Fractals. 2020;139:16. https://doi.org/10.1016/j.chaos.2020.110049.
28. Shi J, Wen Z, Zhong G, Yang H, Wang C, Huang B, Liu R, He X, Shuai L, Sun Z, Zhao Y, Liu P, Liang L, Cui P, Wang J, Zhang X, Guan Y, Tan W, Wu G, Bu Z. Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS-coronavirus 2. Science. 2020;368(6494):1016–20. https://doi.org/10.1126/science.abb7015.
29. Singh RK, Rani M, Bhagavathula AS, Sah R, Rodriguez-Morales AJ, Kalita H, Nanda C, Sharma S, Sharma YD, Rabaan AA, Rahmani J, Kumar P. Prediction of the COVID-19 pandemic for the top 15 affected countries: advanced autoregressive integrated moving average (ARIMA) model. JMIR Public Health Surveill. 2020;6(2): e19115.
30. Tran DA, Bui MD (2018) Long short term memory for monthly rainfall prediction in Camau, VIETNAM.
31. WHO. Modes of transmission of virus causing covid-19 implications- for ipc precaution recommendations. Geneva: WHO; 2020.
32. WHO. WHO Coronavirus disease (COVID-19) dashboard. Geneva: WHO; 2020.
33. WHO. WHO director general's opening remarks at the mission briefing on covid-19. Geneva: WHO; 2020.

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

Authors and Affiliations

Muhammad Shoaib[1] · Hamza Salahudin[2] · Muhammad Hammad[3] · Shakil Ahmad[4] · Alamgir Akhtar Khan[5] · Mudasser Muneer Khan[6] · Muhammad Azhar Inam Baig[7] · Fiaz Ahmad[8] · Muhammad Kaleem Ullah[9]

Hamza Salahudin
hamzasalahudin1@gmail.com

Muhammad Hammad
hammadpattal93@gmail.com

Shakil Ahmad
shakilahmad@nice.nust.edu.pk

Alamgir Akhtar Khan
alamgir.khan@mnsuam.edu.pk

Mudasser Muneer Khan
mudasserkhan@bzu.edu.pk

Muhammad Azhar Inam Baig
azharinam@bzu.edu.pk

Fiaz Ahmad
fiazahmad@bzu.edu.pk

Muhammad Kaleem Ullah
muhammad.kaleem1@ce.uol.edu.pk

[1] Agricultural Engineering Department, Bahauddin Zakariya University, Multan, Pakistan

[2] Agricultural Engineering Department, Bahauddin Zakariya University, Multan, Pakistan

[3] Department of Agricultural Engineering, Bahauddin Zakariya University, Multan, Pakistan

[4] NUST Institute of Civil Engineering, National University of Sciences and Technology, Islamabad, Pakistan

[5] Department of Agricultural Engineering, MNS University of Agriculture, Multan, Pakistan

[6] Department of Civil Engineering, Bahauddin Zakariya University, Multan, Pakistan

[7] Department of Agricultural Engineering, Bahauddin Zakariya University, Multan, Pakistan

[8] Department of Agricultural Engineering, Bahauddin Zakariya University, Multan, Pakistan

[9] Department of Civil Engineering, The University of Lahore, Lahore, Pakistan