Results in Physics 21 (2021) 103817

Contents lists available at ScienceDirect

Results in Physics

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

Forecasting of COVID-19 cases using deep learning models: Is it reliable and practically significant?

Jayanthi Devaraj [a], Rajvikram Madurai Elavarasan [b,*], Rishi Pugazhendhi [c], G.M. Shafiullah [d], Sumathi Ganesan [a], Ajay Kaarthic Jeysree [a], Irfan Ahmad Khan [b], Eklas Hossain [e]

[a] Department of Information Technology, Sri Venkateswara College of Engineering, Chennai 602117, India
[b] Clean and Resilient Energy Systems (CARES) Laboratory, Texas A&M University, Galveston, TX 77553, USA
[c] Department of Mechanical Engineering, Sri Venkateswara College of Engineering, Chennai 602117, India
[d] Discipline of Engineering and Energy, Murdoch University, 90 South St, Murdoch, WA 6150, Australia
[e] Department of Electrical Engineering and Renewable Energy, Oregon Renewable Energy Center (OREC), Oregon Institute of Technology, Klamath Falls, OR 97601, USA

ARTICLE INFO

Keywords:
Artificial Intelligence (AI)
Deep learning
Long short-term memory
Stacked LSTM
ARIMA
Prophet
COVID-19 pandemic
Sustainable Development Goals (SDGs)

ABSTRACT

The ongoing outbreak of the COVID-19 pandemic prevails as an ultimatum to the global economic growth and henceforth, all of society since neither a curing drug nor a preventing vaccine is discovered. The spread of COVID-19 is increasing day by day, imposing human lives and economy at risk. Due to the increased enormity of the number of COVID-19 cases, the role of Artificial Intelligence (AI) is imperative in the current scenario. AI would be a powerful tool to fight against this pandemic outbreak by predicting the number of cases in advance. Deep learning-based time series techniques are considered to predict world-wide COVID-19 cases in advance for short-term and medium-term dependencies with adaptive learning. Initially, the data pre-processing and feature extraction is made with the real world COVID-19 dataset. Subsequently, the prediction of cumulative confirmed, death and recovered global cases are modelled with Auto-Regressive Integrated Moving Average (ARIMA), Long Short-Term Memory (LSTM), Stacked Long Short-Term Memory (SLSTM) and Prophet approaches. For long-term forecasting of COVID-19 cases, multivariate LSTM models is employed. The performance metrics are computed for all the models and the prediction results are subjected to comparative analysis to identify the most reliable model. From the results, it is evident that the Stacked LSTM algorithm yields higher accuracy with an error of less than 2% as compared to the other considered algorithms for the studied performance metrics. Country-specific analysis and city-specific analysis of COVID-19 cases for India and Chennai, respectively, are predicted and analyzed in detail. Also, statistical hypothesis analysis and correlation analysis are done on the COVID-19 datasets by including the features like temperature, rainfall, population, total infected cases, area and population density during the months of May, June, July and August to find out the best suitable model. Further, practical significance of predicting COVID-19 cases is elucidated in terms of assessing pandemic characteristics, scenario planning, optimization of models and supporting Sustainable Development Goals (SDGs).

Introduction

When humans are pioneering in technological progress and simultaneously, dealing with the problem of the climate crisis, a new virus succeeded in infecting humanity. The World Health Organization (WHO) declared the novel coronavirus disease outbreak as a pandemic. The virus transmission is based on the perspective of sources that are infected, susceptibility and viral latency [1]. This disease outbreak would cause serious menaces to human life and society [2,3]. At present, there is no specific treatment for fighting against the pandemic, and various possible antiviral therapies, plasma transfusion, etc., have been cautiously applied in clinical field [4]. The design and identification of new vaccines are important even though old anti-viral drugs are used currently to treat COVID patients [5]. Until effective vaccines are made available, global deaths can only be minimized by suppressing the community transmission and by implementing strict public health measures similar to those developed and implemented during SARS [6]. Human to human transmission can be limited by following certain

* Corresponding author.
E-mail address: rajvikram787@gmail.com (R. Madurai Elavarasan).

https://doi.org/10.1016/j.rinp.2021.103817
Received 16 October 2020; Received in revised form 4 December 2020; Accepted 3 January 2021
Available online 14 January 2021

preventive measures such as washing hands, maintaining social distancing and wearing masks. Besides, it is also the responsibility of the public health authorities to monitor the current status and the outbreak frequencies [7] meanwhile, the public should cooperate to the meaningful measures.

A one health approach can be initiated to decrease the risk of pandemic disease and challenges at the human-animal-environment interface. The multi-disciplinary one health approach can solve the complex problems by utilizing the combined efforts of human and veterinary measures to improve the health of humans and animals [8]. But humans are gifted with the weapon of technology and utilizing them wisely would turn the table against the pandemic. One such technological development that can ultimately be supportive during the pandemic is the forecasting of the infection status ahead. And this work is devoted to analyzing the prediction feasibility and reliability of existing deep learning models. A review of deep learning methods like Generative Adversarial Network, Extreme Machine Learning and Long Short-Term Memory (LSTM) and challenges of AI-based platforms for COVID 19 is discussed. Also, the applications of different types of data in AI-based platforms are discussed in detail [9].

The objective of this study is to perform a comparative analysis on the prediction models such as ARIMA, LSTM, Stacked LSTM and Prophet approaches to predict the growth of COVID-19 concerning the number of infected individuals, the number of deaths and the number of recovered cases. The prediction accuracies are to be compared and, accordingly, the most suitable model is selected based on the various performance metrics and through statistical hypothesis analysis. The approach involves evaluating these models by applying to the global growing cases to check the reliability. Then, country-specific (India) and city-specific (Chennai) predictive analysis are presented as a case study from a real-time forecasting perspective. Further, the role of prediction during the pandemic is analyzed from different perspectives to impart practical significance and to identify what forecasting of infection status means to society.

Section “Literature review” describes the recent literature on the prediction of COVID 19 cases. Section “Deep Learning-based Time Series Forecasting” gives a brief description of the deep learning-based time series prediction models. Section “Forecasting methodology” presents the methodology of the forecast. Meanwhile, the performance analysis of the various prediction models for the COVID-19 infection status is elucidated and a comparative analysis between the models is performed in Section “Results and discussions”. Section “Country and City-specific predictive analysis – a case study” discusses a case study of COVID-19 spread in India and Chennai separately from prediction aspects. Section “Statistical analysis” discusses statistical and correlation analysis. Section “Multivariate stacked LSTM model for COVID 19 prediction” describes the demonstration of multivariate time series data. Section “Unleashing the practical potentiality of prediction during COVID scenario” maps the practical significance of forecasting the infection status in society and finally, conclusions are drawn in Section “Conclusions”.

Literature review

There exist several literature studies that focus to predict the transmission of the COVID-19 virus and to analyze the existing state of spread. The existing literature publications and the research contributions are addressed in this Section.

A mathematical model based on the sequential Monte Carlo simulation was implemented to identify the early transmission rate of the virus by computing daily mean reproduction number Rt with varying parameters like the proportion of cases and confirmed case probability. The outbreak risk can be increased if the transmission is homogeneous [10]. A mathematical model was derived based on the isolation and contact tracing to control the virus transmission. The delay from the onset of symptoms to isolation was determined which increases the probability of spread. There is uncertainty for knowing the symptoms at the early stage and the testing threshold is less which increases the delay and thus, more people are likely to get affected [11]. New technologies like Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), and Big Data can be employed to find different predictions on various aspects to fight against COVID-19. The major areas where the technology can be applied are, for example, early diagnosis of disease, contact tracing, development of drugs, vaccines, predicting the future likely cases, etc. [12,13]. A K-Means clustering algorithm which is an unsupervised machine learning algorithm was used to cluster the COVID-19 data with different variables and concepts for prediction. The model helps in analyzing the countries that are affected and are likely to get affected in the near future [14]. A clustering technology was employed to identify the disease transmission rate in Singapore based on the travel history from China. To reduce the spreading of the virus, the clusters are able to predict local transmission rates that are likely to be affected. However, there is a need for a large volume of data sets in order to develop an adequate model with higher prediction accuracy [15]. AI-based predictive models and their corresponding outcomes are identified in different fields where AI can empower insights into controlling the spread of the pandemic. By training the huge volume of the dataset, the deep learning models can automate the diagnosis, treatment and monitoring of patients which can help health care professionals in various aspects [16,17]. Deep learning Time-series predictions using Recurrent Neural Networks (RNN) are capable of handling non-linearity as well as data dependencies [18,19]. But the modelling fails in capturing the huge dependencies in the time sequences and it is necessary to build a predictive model that captures the non-linear nature of the data with active learning. With data intelligence, the model should be able to assess the probability of pandemic disease. Some of the recent findings on the prediction of COVID 19 cases and the comparison with the proposed work are presented in Table 1.

Research gap and motivation

The spread of the virus should be forecasted accurately for the upcoming weeks and months by analyzing the data in real-time. Mostly, the models like ARIMA, NARNN, SVR, Prophet and Deep Learning models like Deep LSTM/Stacked LSTM, Convolutional LSTM and Bidirectional LSTM have implemented for the prediction of COVID-19 cases. From the literature review, we can infer that a wide range of models exist for time-series predictions with each model excelling in certain conditions and also possessing different limitations. All models are demonstrated for short-term to medium-term predictions. Some of the models provide better results than the other models but provides accurate results only for short term prediction. Besides, the prediction analysis was carried out only for a specific area with limited data. To add upon it, the statistical significance analysis was not carried out to select the best model.

In order to overcome the limitations in the existing system, the proposed work focuses on the analysis of medium-term prediction using ARIMA, LSTM, SLSTM, Prophet models for forecasting global wide COVID-19 cases as well as for country- and city-specific prediction. The

Table 1
Comparison of existing works on COVID-19 prediction methodologies with the proposed work.

Ref.	Forecasting method (Learning Algorithm)	Forecasting horizon	Type of data and Sample size	Data source	Accuracy	Purpose of prediction
Proposed work	Comparative analysis of time series forecasting using ARIMA, LSTM, SLSTM and Prophet	30, 60 and 90 days ahead prediction is done.	Global-wide, country and city specific analysis data from 22nd Jan 2020 to 8th May 2020. Simulated dataset for seven cities for the months of May, June, July and August 2020. All countries data from January 2020 to September 2020.	Datasets were collected from John Hopkins University, World Weather Page and Wikipedia page.	SLSTM outperformed other models. In statistical analysis, ARIMA outperformed LSTM model. Overall, SLSTM model is better than other models.	i. Global-wide, Country specific and city specific cumulative COVID cases prediction is done. ii. Feature correlation is done and best model prediction is identified through statistical hypothesis testing. iii. Multivariate analysis and prediction of India COVID cases is done.
Kırbaş et al. [20]	ARIMA, Nonlinear Autoregression Neural Network (NARNN) and Long-Short Term Memory (LSTM)	14 day ahead forecast	Cumulative confirmed cases data of 8 different European countries and the dataset is considered till 3, May 2020	European Center for Disease Prevention and Control	MAPE values of LSTM model are better than the other models.	To model and predict the cumulative confirmed cases and total increase rate of the countries was analyzed and compared. LSTM outperforms other models.
Arora et al. [21]	Deep LSTM/Stacked LSTM, Convolutional LSTM and Bidirectional LSTM	Daily and weekly predictions	Confirmed cases in India. March 14, 2020 to May 14, 2020	Ministry of Health and Family Welfare	Bi-directional LSTM provides better results than the other models with less error.	Daily and weekly predictions of all states are done to explore the increase of positive cases.
Zeroual et al. [22]	RNN (Recurrent Neural Network), LSTM, Bi-LSTM(Bi-directional), VAE (Variational AutoEncoder)	17 days ahead forecast	Daily confirmed and recovered cases for six countries. Data from 22, January 2020 till 17, June 2020	Center for Systems Science and Engineering (CSSE) at Johns Hopkins University	Based on the performance metrics, VAE outperformed other models in forecasting the pandemic.	To forecast the number of new COVID-19 cases and recovered cases.
Shahid et al. [23]	ARIMA, support vector regression (SVR), long short-term memory (LSTM), Bi-LSTM	48 days ahead forecast	22 January 2020 to 27 June 2020. 158 samples of the number of confirmed cases, deaths and recovered cases.	Dataset is taken from the Harvard University	Bi-LSTM outperforms other models with lower R^2 score values.	To predict the number of confirmed, death and recovered cases in ten countries for better planning and management.
Chimmula and Zhang [24]	LSTM	14 days ahead forecast	confirmed cases of Canada and Italy till 31, March 2020	Johns Hopkins University and Canadian Health authority	92% accuracy	To predict the number of confirmed cases of Canada and Italy and to compare the growth rate.
Alzahrani et al. [25]	ARIMA, Autoregressive Moving Average (ARMA)	1 month ahead forecast	Cumulative daily cases from March 2, 2020, to April 20, 2020	Daily and cumulative confirmed COVID-19 cases in Saudi Arabia were collected from Saudi Arabia Government website.	ARIMA performs well than ARMA, MA and AR.	To predict the daily reproduction of confirmed cases one month ahead.
Ogundokun et al. [26]	Linear regression model	8 days ahead forecast	March 31, 2020 to May 29, 2020	NCDC website	95% confidence interval	To predict the COVID-19 confirmed cases in Nigeria.
Ribeiro et al. [27]	ARIMA, cubist regression (CUBIST), random forest (RF), ridge regression (RIDGE), support vector regression (SVR), and stacking-ensemble learning	1,3 and 6 days ahead forecast	Cumulative confirmed cases in Brazil until April, 18 or 19 of 2020	The dataset was collected from an application programming interface that retrieves the daily data about COVID-19 cases which are publicly available	Based on the performance metrics, SVR, and stacking-ensemble learning outperformed other models	To predict the cumulative confirmed cases in Brazil
Tomar and Gupta [28]	LSTM	30 days ahead forecast	Cumulative and daily dataset of COVID-19 cases in India	Center for Systems Science and Engineering (CSSE) at Johns Hopkins University	LSTM has got 90% accuracy in predicting COVID cases	To predict the number of confirmed and recovered cases using data-driven estimation method.
Car et al. [29]	Multilayer Perceptron (MLP) artificial neural network (ANN)	30 days ahead forecast	22nd January 2020 to 12th March 2020 Infected, recovered and deceased data	Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) and supported by ESRI Living Atlas Team and the Johns	Higher accuracy for confirmed cases with 0.986R^2 Value	To predict the spread of pandemic world-wide.

(*continued on next page*)

Table 1 (*continued*)

Ref.	Forecasting method (Learning Algorithm)	Forecasting horizon	Type of data and Sample size	Data source	Accuracy	Purpose of prediction
				Hopkins University Applied Physics Lab (JHU APL)		
Shastri et al. [30]	LSTM, Stacked LSTM, Bi-directional LSTM and Convolutional LSTM	30 days ahead forecast	India and USA-Confirmed cases data from 7th Feb to 7th July 2020 Death cases data from 12th March to 7th July 2020.	Datasets of India and USA are taken from the Ministry of Health and Family Welfare, Government of India and Centers for Disease Control and Prevention, U.S Department of Health and Human Services.	ConvLSTM outperforms stacked and bi-directional LSTM in confirmed and death cases.	To predict the COVID-19 confirmed and death cases one month ahead and to compare the accuracy of deep learning models
Hawas [31]	Recurrent Neural Network (RNN)	30 days and 40 days ahead forecast	Daily confirmed cases in Brazil 54 to 84 days 7th April to 29th June 2020	Center for Systems Science and Engineering (CSSE) at Johns Hopkins University	Achieved 60.17% accuracy.	To predict one month ahead confirmed cases and to take preventive measures.
Papastefanopoulos et al. [32]	Six different forecasting methods are presented. ARIMA, the Holt-Winters additive model (HWAAS), TBAT, Facebook's Prophet, Deep AR	7 days ahead for the ten countries	Jan 2020 to April 2020 and the population of countries.	Novel Corona Virus 2019 Dataset and population-by-country dataset from kaggle.com	ARIMA and TBAT outperformed other models in forecasting the pandemic	To predict the future COVID-19 confirmed, death and recovered cases by considering the country population.

sequential data prediction is done each having unique characteristics by considering all countries data. Thus, a comparative analysis for predicting COVID-19 infected cases would yield the best model for forecasting accurate data for the defined conditions. To the best of our knowledge, there is no study to predict all the essence of infection details like confirmed, death and recovered cases for medium-term prediction in various scales such as global-wide, country-wide and city-specific, especially by using the above-mentioned deep learning-based time series prediction models. The proposed work also involves correlation analysis to find out the relationship of growth rate with other external factors like temperature, rainfall etc. by considering the monthly data. Statistical hypothesis analysis is also applied to determine the best suitable model. Analysis of Multivariate time series prediction using LSTM is done to predict the increase in the number of infected cases.

In summary, the novelty of our work is presented below in the consequent points.

- Prediction of confirmed, death and recovered cases using various deep learning models for world-wide analysis are carried out and the comparison of the performance is also accomplished.
- Sorting the most reliable model from the statistical analysis.
- Multivariate stacked LSTM for long-term prediction is done.
- A descriptive case study of the country and city-specific analysis of India and Chennai are analyzed in detail.
- Revealing the prediction potentiality under pandemic scenario – from pandemic assessment, scenario planning, effective optimization and SDGs perspective.

Deep Learning-based time series forecasting

Deep Learning (DL) gives promising results in time series data analysis and forecasting. DL models are capable of learning the temporal dependencies and structures such as trends and seasonality in the data automatically. Multi-Layer Perceptron (MLP) [33] can handle multivariate inputs and can be used for multi-step forecasting. Feed Forward Neural Networks (FFNNs) [34] with sparse representation can be useful for time series prediction. Convolutional Neural Networks (CNNs) [35] are used for automatic feature learning and can support multivariate inputs and outputs for time series forecasting. ML and DL models play a major role in accurately predicting the progression of diseases. An ensemble approach of Support Vector Machine (SVM), Neural Networks (NNs) and Naive Bayes was used to predict the disease risk and the condition of a patient one day in advance by training the past k days historical medical measurements [36]. It is important to extract the features of data with high dimensionality, without error and noise. Deep features are extracted using Recurrent Neural Networks (RNNs) with a de-noising auto-encoder which can effectively encode patients' hospital records for mortality and comorbidity prediction [37]. The ubiquitous nature of multivariate data can be handled by RNNs to predict the temporal dependencies in the time series data and these can also identify the missing patterns in the data which improves the prediction accuracy [38]. This section describes the necessary concepts of time series forecasting models like ARIMA, LSTM, SLSTM and Prophet used in this study for forecasting the pandemic COVID-19 outcomes.

ARIMA

ARIMA is a time series forecasting model which is a form of regression analysis and is used to predict the future trends on the time series dataset. This model is used to capture the autocorrelation from the data which computes the future values based on the correlations between the previous values. The average of the error term is zero and the variance is expressed as σ^2. If Y_t denotes the time series value at time t, then the p-order autoregressive process expression is represented as in Eq. (1) and is shown as AR(p).

$$Y_t = \delta + \varnothing_1 Y_{(t-1)} + \varnothing_2 Y_{(t-2)} + \cdots + \varnothing_P Y_{(t-p)} + \in_t \quad (1)$$

Here, δ is a constant value and $\in_t$ is the error term. The q^{th} degree of moving average process MA(q) is represented in Eq. (2).

$$Y_t = \mu + \in_t + \theta_1 \in_{t-1} + \theta_2 \in_{t-2} + \cdots + \theta_q \in_{t-q} \quad (2)$$

An Auto Correlation Function (ACF) plot is used to visualize the correlations between the data points where the x-axis represents the correlation coefficient and y-axis represents the lag units [20]. By combining two AR(p) and MA(q) equations, the expression of ARIMA(p, q) can be obtained and is given in Eq. (3).

$$Y_t = \delta + \varnothing_1 Y_{t-1} + \cdots + \varnothing_p Y_{t-p} + \in_t + \theta_1 \in_{t-1} + \cdots + \theta_q \in_{t-q} \quad (3)$$

If the processed time series is not stationary, it can be made stationary by taking the difference between process d times (ΔY_t). Eq. (4) denotes the ARIMA (p, d, q)$_n$ process.

$$\left(1 - \varnothing_p \mathrm{L} - \varnothing_1 \mathrm{L}^2 - \ldots - \varnothing_p \mathrm{L}^q\right) \Delta^d \mathrm{Y}_t = \delta + \in_t + \theta_1 \in_{t-1} + \ldots + \theta_q \in_{t-q} \quad (4)$$

The likelihood of the data is denoted as L, the lag operator, and p, q and m denotes the order of the autoregressive part, the order of the moving average part and intercept of the model, respectively. According to these parameters, the model with lowest Akaike Information Criterion (AIC) is considered as good than the others to determine the order of non-seasonal ARIMA model. First, it is necessary to transform the degree of non-stationary to stationary series and the order of differencing is identified. The data set is divided into train and test data to validate the accuracy of the model. n is the number of periods to forecast which is set before building a model. The model can be validated by comparing the predicted and actual values.

Long short-term memory (LSTM) and stacked LSTM

Complex relationships can be handled by Artificial Neural Networks (ANNs) but they are not capable of capturing historical dependencies in the data and the forecasting accuracy mainly depends on the features of the dataset [39]. A type of ANN called a Recurrent Neural Network (RNN) can handle temporal dependencies in the data using network loops. In RNN, the current state is predicted based on the previously hidden state values and the value of the current input. This is employed for solving problems involving sequential decision making [40]. However, RNN is used for short term forecasting and it cannot predict the long-term dependencies from the data.

LSTM networks overcome the drawback of RNN with the memory cells, input gate, forget gate and output gate in the network for efficient sequence prediction.

Fig. 1 shows the architecture of the LSTM in which the sigmoid layer is used by the forget gate to determine the state to be preserved. Data moves through the components known as cell states. LSTM contains memory blocks which contain hidden units that are used to control the flow of information from input to output ports. The first sigmoid function is the forget gate which forgets the previous cell state information. The input gate denotes the next sigmoid and the first tanh function which indicates the information saved to the cell state or what information should be forgotten. The last sigmoid function is the output gate which determines the information to be passed to the next hidden state. Weights are adjusted using the input gate i_t, forget gate f_t, and output gate o_t.The input gate determines the value of c_t and c_{t-1}, the information to be sent at timestamps t and t-1with x_t as an input vector. Using the sigmoid layer and tanh layer, the output of the LSTM cell is tuned by the output gate o_t and h_t represents the hidden state at timestamp t. Forget gate determines which data should be removed that are not relevant to the past timestamp values. W_f, $W_c$and W_0 represents the input weights, $b_f, b_i, b_c, and\ b_0$ represents bias weights and Θ shows the point-wise multiplication of vectors which are represented mathematically in Eqs. (5)–(10) as below:

$$f_t = \sigma\left(W_f \cdot [x_t \cdot h_{t-1}] + b_f\right) \quad (5)$$

$$i_t = \sigma\left(W_f \cdot [h_{t-1} \cdot x_t] + b_i\right) \quad (6)$$

$$c_{i'} = tanh(W_c \cdot [h_{t-1} \cdot x_t] + b_c) \quad (7)$$

$$c_t = f_t \Theta c_{t-1} \oplus i_t \Theta c_{i'} \quad (8)$$

$$o_t = \sigma(W_0 \cdot [h_{t-1} \cdot x_t] + b_0) \quad (9)$$

$$h_t = o_t \Theta tanh(c_t) \quad (10)$$

The performance of the LSTM is highly dependent on selecting the hyper-parameters for achieving good results [41]. Long Short-Term Memory is used for handling the sequences in the input observations and is capable of learning mapping of input to output functions which are not supported by MLP and CNN.

Stacked/Deep LSTM

Stacked LSTM/Deep LSTM consists of more hidden LSTM layers with multiple memory cells in each layer. LSTM is used for forecasting sequence prediction problems with time series data which produces output per single input step rather than producing single output for all time steps. In Stacked LSTM, multiple LSTM layers are stacked together to make an accurate model with high level deeper representation [42]. The previous layers' representations are learned by the next higher layers for better optimization.

Fig. 2 shows the sequence of LSTM layers in which the current layer that receives the value from the previous layers produces a higher level of abstraction with more complexity and representation. In the proposed work, LSTM and Stacked LSTM are implemented using a Keras package where the number of hidden layers can be added to the sequential model and the value of the return sequence attribute should be set to true for modelling.

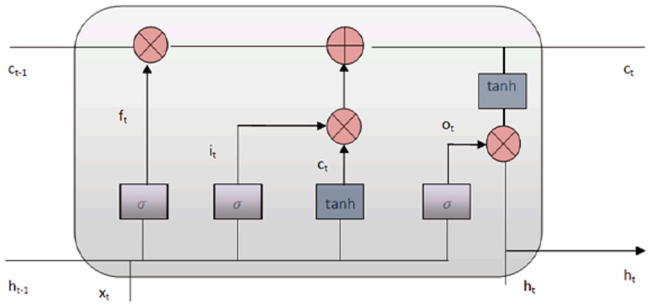

Fig. 1. The architecture of LSTM.

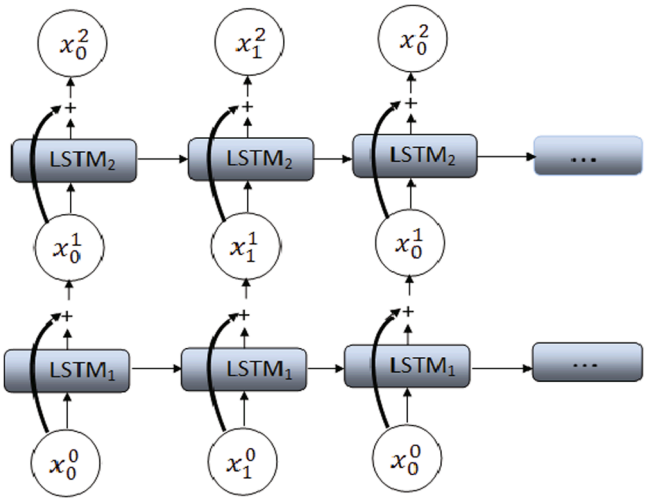

Fig. 2. The architecture of Stacked LSTM.

Prophet model

The Prophet model is used for forecasting time series data in the medium-term and long-term. The three main decomposed components of time series models are trends, seasonality and holidays. These components can be combined as specified in the Eq. (11) as below.

$$y(t) = g(t) + s(t) + h(t) + \varepsilon t \tag{11}$$

- g(t): non-periodic changes in time series data are modelled using a logistic growth curve.
- s(t): periodic changes in the data are modelled (e.g., weekly/yearly seasonality).
- h(t): captures holiday effects with irregular schedules.
- εt: any abnormal changes accommodated by the model are considered.

Time is used as the regressor which tries to fit linear and non-linear functions as additive components for modelling [43]. A factor that multiplies g(t) as a seasonal effect is obtained through a log transform and is called multiplicative seasonality.

Trend, seasonality and other events

Trend modelling involves fitting a piecewise linear curve or a non-linear saturating growth model. Growth forecasting is used to analyze how the COVID-19 cases have grown so far, and how they are likely to continue growing in the near future. This can be determined using the logistic growth model which is represented in Eq. (12) as below.

$$g(t) = C/(1 + exp(-k(t-m))) \tag{12}$$

where C indicates the growing capacity, k specifies the growth rate and m is an offset parameter. Both the carrying capacity and rate of growth are not constant. The model can fit the historical data at varying rates. Change points are explicitly defined which allows the growth rate to change. The generative model is used to forecast the uncertainty in the COVID-19 trend. By altering the parameter rate, the flexibility of the model can be controlled.

Prophet uses Fourier series to forecast the seasonality effects and the seasonality models are specified as the periodic functions of t [44]. The arbitrary smoothing of seasonal effects with a scaling time variable using Fourier series is represented as

$$s(t) = \sum_{n=1}^{\infty} \left(a_n \cos \frac{2n\pi t}{P} + b_n \sin \frac{2n\pi t}{P} \right) \tag{13}$$

where P is the period and, for a given value of N, to fit the seasonality model the parameters a_1, a_2,……..,a_n and b_1,b_2,...,b_n need to be estimated. The model selection procedure can be used to automate the selection and tuning of the parameters. The value of N can be tuned for better accuracy [45].

All the four models discussed above are used to forecast the COVID-19 confirmed, recovered and death cases in the near future.

Forecasting methodology

This section presents the step by step methodology needed to develop the prediction models to forecast the future COVID-19 cases. The model requires COVID-19 data such as confirmed cases, recovered cases and death cases as input to forecast the corresponding future data for a defined period of time. From the acquired data, data cleaning and a normalization process is carried out to remove the unwanted fields. Feature extraction is done by identifying the dependent and independent variables from the data. The proposed models actively learn real-time data with current observations of COVID-19 in order to predict future outbreaks.

Python being a high-level general-purpose programming language, it is used to interact with deep learning libraries as application program interfaces (APIs). The experiments are carried out using open source libraries such as NumPy, Pandas, TensorFlow (Google) and Keras (Deep Learning Framework). Fig. 3 shows the typical architecture of the proposed model to predict the future count of confirmed, recovered and death cases. The comparative analysis of different models like ARIMA, LSTM, Stacked LSTM and Prophet are done and the best prediction model is identified based on the prediction results.

The overall methodology involves the following process. a. Data Collection and Data Simulation, b. Data Preprocessing, c. Prediction of COVID 19 cases using various models, d. Correlation analysis and statistical hypothesis testing to select the suitable model, e. Model training, testing and evaluation, f. Analysis of multivariate LSTM for predicting COVID 19 cases.

COVID-19 data collection

There are various publicly available data sources released by governments and the real-time observations are being included for up-to-date analysis for the prediction of COVID-19 outcomes. Three types of datasets are used,

a. Time series prediction of global cases - The global dataset is being collected from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, USA [46]. Three different time-series datasets such as confirmed, recovered and death cases are collected for analysis. It contains variables like name of province, country, latitude, longitude and number of cases with respect to date. The number of confirmed, recovered and death cases by country is provided starting from 22/1/2020 (DD/MM/YYYY) to 08/05/2020 for each type of cases. The data is transformed into the number of confirmed, death and recovered cases per day with the global data

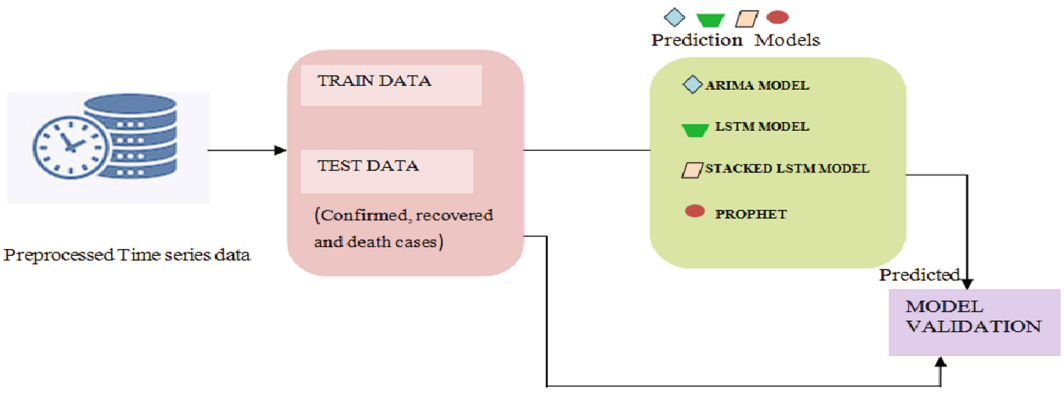

Fig. 3. Architecture diagram of predictive model infected cases.

and the few processed records are shown in the analysis of respective models. With the data set acquired, the prediction of the total number of cumulative cases in the short-term and medium-term all over the world is accomplished.

b. Simulated dataset for correlation analysis - Simulated dataset with 16 features such as average temperature values for the months of May, June, July, August and average rainfall values for the months of May, June, July, August, population, area, population density, city, total infected cases for May, June, July and August. The external factors data was collected from the world weather page. The simulated dataset was created for seven different cities in Tamil Nadu. Other features like population, area and population density are obtained from the Government website.

c. Combined time series dataset for multivariate analysis – Multivariate stacked LSTM prediction is analyzed with the combined dataset of multiple features like date, country/region, province/state, latitude, longitude, confirmed, death and recovered cases with 80,970 records from 22/01/2020 to 17/11/2020. Dataset is being collected by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, USA [46].

Real-world datasets collected may be inconsistent and analyzing the raw data may lead to erroneous results. Hence, data needs to be preprocessed for analysis. To ensure consistencies of knowledge discovery data, there are various preprocessing techniques available to deal with messy data [47]. The attributes of multiple files can be combined to create a single file in a usable format [48]. Transformation involves scaling of attributes for further analysis. The number of attributes can be reduced by removing the redundancies present in the dataset using data reduction strategies [49,50]. MinMax scaler is used to normalize the data to avoid bias during the training of data.

Performance metrics and model window size

The performance of the models was evaluated based on the predicted outcome values using common statistical measures [51]. In this study, the evaluation metrics of Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Correlation Coefficient (CC) are computed for the forecasting models where $\widehat{y}_i$represents the predicted output and y_i represents the actual output which is shown in Eqs. (14)–(19).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}(|\widehat{y}_i - y_i|) \tag{14}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{(|\widehat{y}_i - y_i|)}{y_i}\right)*100\% \tag{15}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y}_i - y_i)^2 \tag{16}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{y}_i - y_i)^2} \tag{17}$$

$$CC = \frac{\frac{1}{n-1}\sum_{i=1}^{N}\left(y_i - \widehat{y}_{ii}\right)^2\left(y_i^* - \widehat{y}_{ii}^*\right)}{\sigma_{y_i}\sigma_{y_i^*}} \tag{18}$$

R^2 is a statistical measure of the fitness of the predicted values to the actual values. It indicates how much variance is explained by the model. If the R^2 value is 0.5, then the model can capture half of the observed variation.

$$R^2 = \frac{\frac{1}{n}\sum_{i=1}^{N}\left(y_i - \widehat{y}_{ii}\right)^2}{\frac{1}{n}\sum_{i=1}^{N}(y_i - y_i^{'})^2} \tag{19}$$

ARIMA, LSTM, Stacked LSTM and Prophet techniques are used to predict COVID-19 cases using time series data. To increase the robustness and flexibility of the model, time frame of 60 days and 30 days ahead from 9th May 2020 has been considered. The prediction results of all the four models are discussed in detail in the subsequent sections.

Results and discussions

The results of the predictions and performance analysis together with comparative analysis is discussed in this Section.

ARIMA model

The prediction results of future COVID-19 cases using the ARIMA model are described in this Section. The time frame for prediction involves 60 days (until end of June 2020) from 9th May 2020 as the starting period for forecasting. The last ten records from the original dataset are considered as the test data starting from 29th April 2020 until 8th May 2020. The prediction accuracy of the ARIMA model with MAPE values are 0.37, 0.74 and 1.12 for confirmed, death and recovered cases respectively. The forecasted values of the confirmed, death and recovered cases from 9th May 2020 are shown in Fig. 4.

Fig. 4(a)–(c) shows that the number of confirmed, death and recovered cases as on May 31st 2020 are predicted to be 6,218,889, 405,840 and 2,485,323, respectively. Also, it is shown that the number of confirmed, death and recovered cases as on June 30th, 2020 are 9,493,908, 575,178 and 4,497,864, respectively. In Fig. 4(a)–(c) the grey area indicates the uncertainty intervals which shows the real observation within the range. It is a useful indicator to measure overfitting. It depicts whether how the future trend changes based on the previous history.

Performance evaluation of the ARIMA model

Table 2 represents the coefficient, standard errors and normal distribution of values for confirmed, recovered and death cases respectively. Auto-Regressive (AR) term is dependent on its own lags with positive autocorrelation at lag 1 and it measures under-differenced series in the forecasting equation. Moving Average (MA) term is dependent on the lagged forecast errors with negative autocorrelation at lag 1 and it measures over-differenced series in the forecasting equation. ARIMA (1,1,1) represents a model with one AR term, a first-order difference and with one MA term applied to the z variable which shows the linear trend in the data. The ar.L1 and ma.L1 represents one autoregressive lag and one moving average lag and sigma 2 represents the variance of the error term. In ACF (Auto-correlation function), the coefficient of correlation is presented in x-axis and the number of lags is represented in the y-axis.

The standardized residual error and the correlogram are represented over time. The Akaike Information Criteria (AIC) measure is used to identify the best fit of the model which is shown in Eq. (20).

$$\mathrm{AIC} = -2\log(\mathrm{L}) + 2(\mathrm{p}+\mathrm{q}+\mathrm{m}) \tag{20}$$

The Correlogram represents the low correlation among the time series residuals and the Normal Q-Q plot shows the distribution of residuals which are approximately normal. By summing up the correlation coefficients, the intercept and the best fit can be identified which shows the weight of each feature. The $p > |z|$ column represents the feature weight significance as described in Table 2. The ARIMA(1,1,1)x (1,1,1,12) model has the lowest AIC value of 1646.45. The Auto Regression (AR) coefficient is estimated to be 0.99 with a standard error of 0.016 and the Moving Average (MA) coefficient is −0.7541 with the

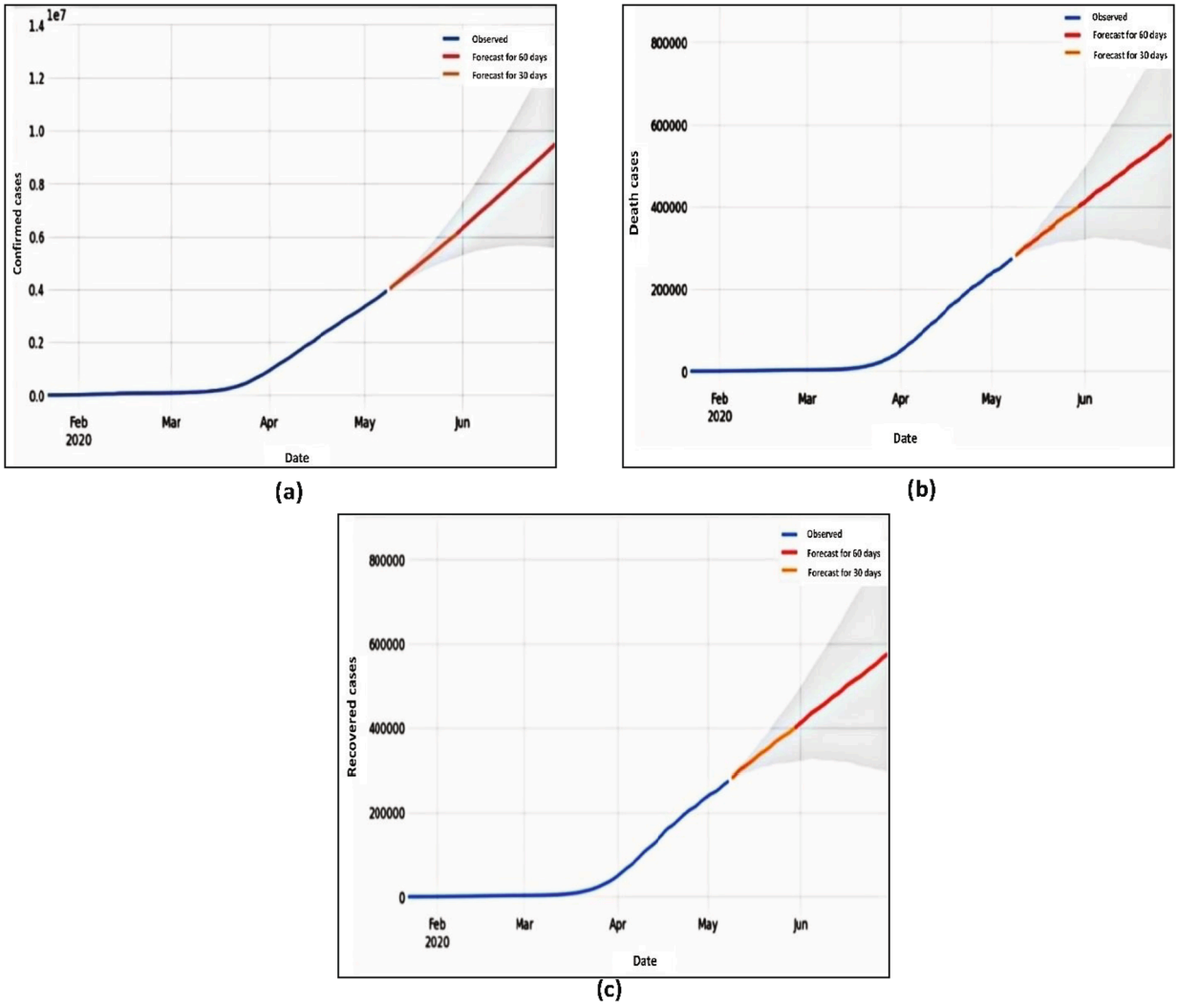

Fig. 4. Prediction of (a) Confirmed, (b) deaths and (c) recovered cases using the ARIMA model.

Table 2
Coefficient and error values for forecasted infected cases using the ARIMA model.

(a) Coefficient and error values of confirmed cases using the ARIMA model

	Coef	std err	z	p>|z|	[0.025	0.975]
ar.L1	0.9715	0.032	30.236	0	0.908	1.034
ma.L1	−0.0876	0.159	−0.552	0.581	−0.398	0.233
ar.S.L12	−0.5586	0.114	−4.892	0	−0.782	−0.335
sigma2	7.60E + 06	2.08E-10	3.65E + 17	0	7.60E + 07	7.60E + 07

(b) Coefficient and error values of death cases using the ARIMA model

	Coef	**std err**	**z**	**p>|z|**	**[0.025**	**0.975]**
ar.L1	0.9214	0.036	25.694	0	0.851	0.992
ma.L1	0.0569	0.081	0.699	0.484	−0.103	0.216
ar.S.L12	−0.6301	0.105	−6.001	0	−0.836	−0.424
sigma2	9.02E + 07	9.67E + 04	9.333	0	7.13E + 05	1.09E + 06

(c) Coefficient and error values of recovered cases using the ARIMA model

	Coef	**std err**	**z**	**p>|z|**	**[0.025**	**0.975]**
ar.L1	0.9992	0.016	64.128	0	0.969	1.03
ma.L1	−0.7541	0.094	−7.991	0	−0.939	−0.569
ar.S.L12	−0.7164	0.173	−4.148	0	−1.055	−0.378
sigma2	5.23E + 07	2.87E-09	1.82E + 16	0	5.23E + 07	5.23E + 07

standard error of 0.094. The lower AIC value of 1334.9, with ARIMA (1,1,1)x(1,1,1,12) is identified as the best model for COVID-19 death cases. The AR coefficient is estimated to be 0.9214 with a standard error of 0.036 and the MA coefficient is 0.0569 with the standard error of 0.081 and the QQ plot represents a normal distribution of values. The lower AIC value of 1647.95, with ARIMA(1,1,1)x(1,1,1,12) is identified as the best model for COVID-19 recovered cases. The AR coefficient is estimated to be 0.9992 with a standard error of 0.016 and the MA coefficient is −0.7541 with the standard error of 0.094.

The ARIMA model gives reasonably good predictions with the MAPE values of 0.372, 0.742, 1.12 for confirmed, recovered and death cases, respectively.

LSTM and stacked LSTM model

LSTM selects the dependent features that have an impact on training from the original dataset. Confirmed cases, death cases and recovered cases are the dependent features of the COVID-19 dataset for training the model. The output is a vector demonstrating the predicted values. The features are converted into a machine-readable format and variable input shapes are handled using the Keras package. The original data is being transformed as the date-wise total number of COVID-19 confirmed, deaths and recovered cases all over the world. All the lists and arrays should be combined with the same shape for further analysis and python lists consume more memory than the numpy arrays.

Selecting the nodes, layers and hyper-parameters

In Keras, it is necessary to initialize the model as Sequential() and multiple layers are stacked one above the other. Based on the trial and error approach of selecting nodes and layers, good prediction results can be obtained depending on the dataset. N_k number of nodes are chosen for testing to estimate the loss as shown in Eq. (21) [52].

$$N_k = \frac{N_s}{(\alpha^*(N_i + N_0))} \tag{21}$$

The above equation represents the input neurons *Ni*, the output neurons *No*, the number of samples in training data N_s, and **α** denotes a scaling factor between 2 and 10. Based on the loss, the optimal model can be identified. The data is normalized and reshaped in the range of (−1,1). Fig. 5(a)–(c) shows the plot of epochs and loss in LSTM training where the x-axis represents the number of epochs and the y-axis represents the learning curve and loss with the scaling of (0,1). The original dataset is divided into training data and test data. The number of hidden LSTM layers are chosen as two which is sufficient to discover complex patterns. The dropout layer helps to ignore some neurons during the training process in order to prevent over-fitting. The value of the dropout layer is 0.2 which is added after every layer of LSTM which retains the accuracy of the model. The parameter settings used are learning rate with the value of 0.0005, the number of steps is 3, the number of features is 01 and the number of hidden units is 100. The density layer is used after all stacked LSTM layers to interpret the output values and ReLu activation function is used. It is clearly shown that there is a good learning rate in plotting loss across epochs. The sequence of data is transformed into an appropriate format for analysis. Every training sample contains a sequence of data points. In order to predict the COVID-19 cases in the future, the sequence of data is passed to the LSTM layers and the output of the last time step is passed to the next input sequence.

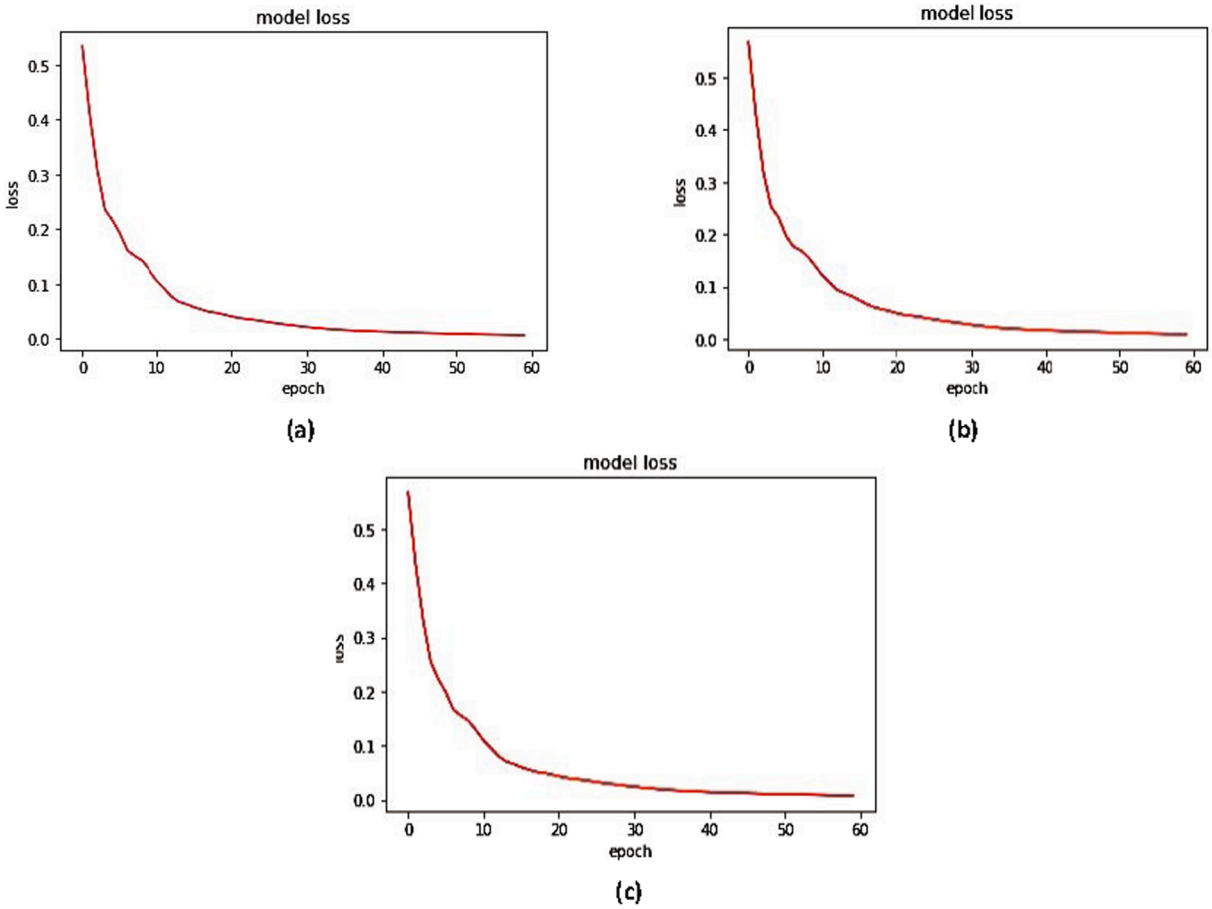

Fig. 5. (a) LSTM loss vs epochs for Confirmed cases (b) LSTM loss vs epochs for Death cases (c) LSTM loss vs epochs for Recovered cases.

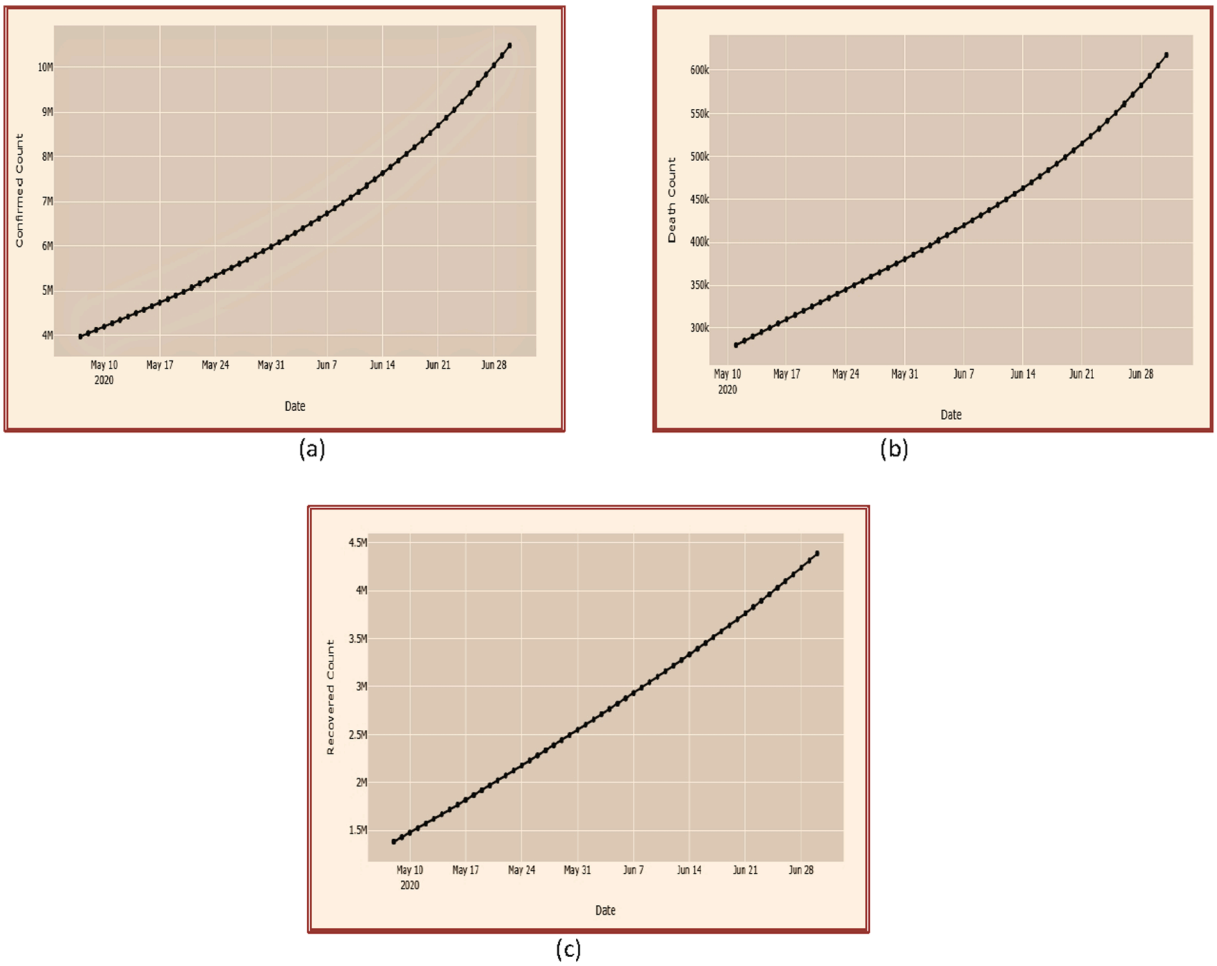

Fig. 6. A plot of (a) confirmed (b) death and (c) recovered cases using LSTM model.

Building and training LSTM and Stacked LSTM models

Adam optimizer, an adaptive optimization algorithm, is used for optimizing the mean square loss and requires minimum tuning of hyper-parameters. The prediction results of future COVID-19 cases using LSTM and Stacked LSTM models are described in this Section. The dataset is divided into training and testing, and the predicted values and the observed values of the test dataset is presented.

Good prediction accuracy is achieved with the predicted confirmed, death and recovered cases for LSTM and SLSTM models. For the LSTM model, MAPE values of confirmed, death and recovered cases are 0.37, 0.53 and 1.07 respectively. For the SLSTM model, MAPE values of confirmed, death and recovered cases are 0.2, 0.43 and 0.9, respectively. For example, on 8th May, the observed value of confirmed, recovered and the death cases are 3,938,064, 274,898 and 1,322,050, respectively. The predicted values using the LSTM model are 3,860,548, 270,470 and 1,317,394 while using the Stacked LSTM model, the predicted values are 3,860,458, 272,631 and 1,335,877 for confirmed, death and recovered cases, respectively, which represents good prediction accuracy.

Prediction of future cases from 9th May 2020

The predicted values of the confirmed, death and recovered cases for the next 60 days starting from 9th May 2020 until 30th June 2020 are depicted in Fig. 6.

Fig. 6(a)–(c) shows the forecast values of confirmed, death and recovered cases using the LSTM model. From the above figures, it can be inferred that the total number of confirmed, death and recovered cases are around 6.2 million, 372 thousand and 2.5 million at the end of May 2020 and 9.4 million, 575 thousand and 4.3 million cases at the end of June 2020, respectively, using the LSTM model.

Fig. 7(a)–(c) shows the forecast values of confirmed, death and recovered cases using the Stacked LSTM model. The blue line indicates the prediction for the month of May 2020 and red line indicates the prediction for the month of June 2020. From the above figures, we can infer that the total number of confirmed, death and recovered cases are around 6.3 million, 380 thousand and 2.9 million at the end of May 2020 and 9.9 million, 580 thousand and 4.9 million cases at the end of June 2020, respectively, using the Stacked LSTM model.

From the above insights, it is shown that the Stacked LSTM model performs well in predicting COVID-19 confirmed/recovered and death cases using the time series dataset with a good learning rate and better accuracy as compared to LSTM.

Prophet model

A data frame with time information in one column and the metric need to forecast in another column is given as input to the Prophet model. After the initialization of the cleaned data frame, a duplicate of the same is created for further analysis. The Prophet model follows a strict condition that the input columns should be named as ds (date stamp) and y (numeric measurement) components which represent time and metric respectively.

Creating future data frame or the prediction of confirmed, death and recovered cases

After initializing the Prophet model, a fit method is invoked with the Data Frame as input. A new data frame for the time series forecast with

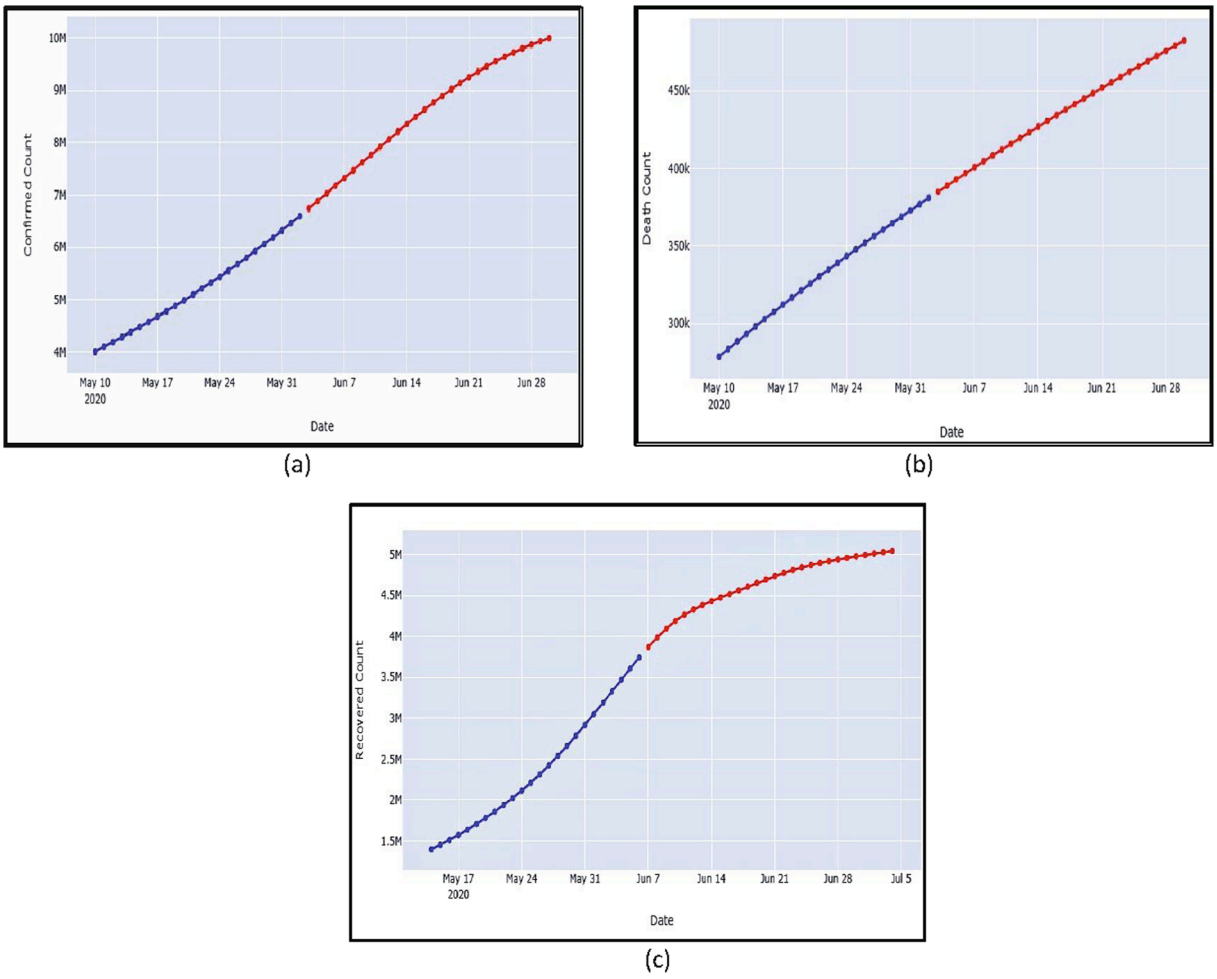

Fig. 7. Plot of (a) confirmed (b) death and (c) recovered cases using Stacked LSTM model.

ds component that specifies the dates we want to make predictions at is created.

Prediction of future confirmed, death and recovered cases

The frequency of time series data should be taken into account while modelling. The inputs to the predict function of the fitted model are the future dates of the DataFrame. Prophet returns the DataFrame with the values of columns as ds and yhat where ds is the date, and yhat is the actual forecast or predicted value of the metric column. The lower bound and upper bound values are represented by yhat_lower and yhat_upper, respectively. Markov Chain Monte Carlo methods are used by the model to generate forecast values.

Prediction of confirmed, death and recovered cases (60 days ahead)

The predicted values of COVID-19 cases until the end of June 2020 are described below. 60 days ahead prediction results for confirmed, death and recovered cases are plotted in Fig. 8.

Fig. 8(a)–(c) presents the expected number of confirmed, death and recovered cases as around 8.15 million, 582,581 and 3.18 million, respectively. The Prophet model prediction on recovered cases is not reasonably good since the number of recovered cases is only 3.18 million as at the end of June 2020, whereas, in the ARIMA, LSTM and SLSTM modelling, the predicted number of recovered cases as at 30th June 2020 are 4.4, 4.3 and 4.9 million, respectively.

Fig. 9(a)–(c) represents the plot of forecast values of confirmed, death and recovered cases and the corresponding additive components. The trend component shows the increase in one month ahead and weekly component shows the daily cases and the trend of increase in the spread of disease. Tuning of multiple regressors can be done for better performance. Error increases if the horizon value of the Prophet model is large for confirmed/recovered/death cases. For the Prophet model, the MAPE values of confirmed, death and recovered cases are 0.39, 0.70 and 1.2, respectively. Grey areas in Figs. 8 and 9(a)–(c) indicates the future trend changes similar to the input data and shows whether the trend changes move forward or upward. It is the uncertainty interval with the default value of 80%.

Comparative analysis

The comparisons of models in terms of considered metrics are highlighted below in Tables 3 and 4.

Table 3 and Fig. 10 indicate that the Stacked LSTM model outperforms all the other three models with lower RMSE and MAPE values and LSTM outperforms ARIMA and Prophet in predicting confirmed and death cases. Based on the evaluation metric values, SLSTM is better than the other models.

The results show that the Stacked LSTM model is the most accurate model for predicting future COVID-19 confirmed, recovered and death cases. A prediction comparison at various time period for the four models is represented in Fig. 11. Table 4 shows the comparison of actual and forecasted values of all the models.

From the results, it is clear that SLSTM has a minimal error in percentage values for confirmed, deaths and recovered cases than the other models.

In the future, the scale chart trend shows that the number of confirmed cases all over the world will be increasing exponentially and

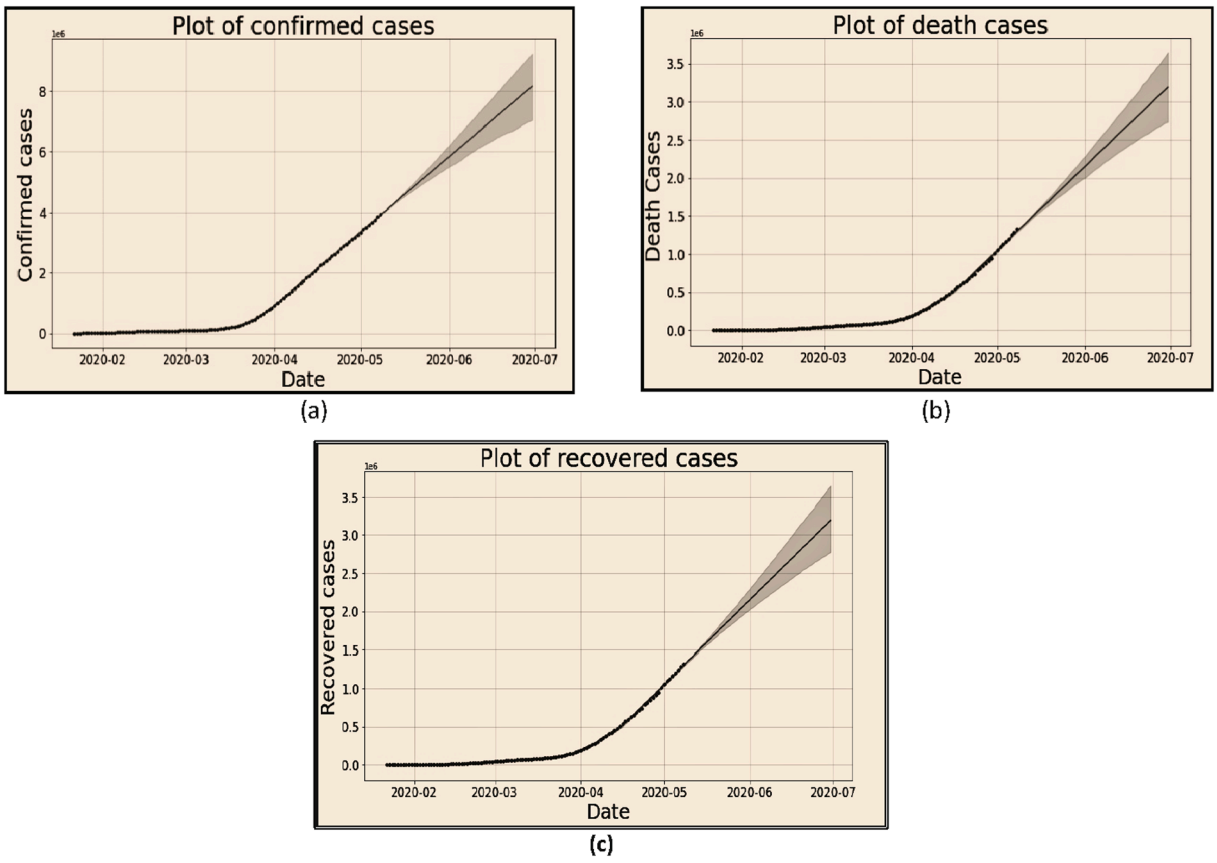

Fig. 8. A plot of (a) confirmed (b) death and (c) recovered cases using Prophet Model.

can be modelled by using Eq. (22) [52].

$$N_d = (1 + E^*P)^d N_0 \tag{22}$$

where the expected future confirmed cases is N_d, infection being detected every day by infected people is E, P is the probability of exposure that leads to COVID-19, the initial number of cases is N_0, d is the days' interval between present given time and future time. When E or P decreases, N_d will decrease. The daily increase rate of confirmed cases and the number of newly infected cases are proportional to the existing cases. The growth factor of COVID-19 depends on the number of newly confirmed cases on a current day and the number of newly confirmed cases on the previous day.

Fig. 12 shows the predicted COVID-19 cases and Fig. 13(a) and (b) represents the error in recovered cases if the training data size increases. The training dataset is taken till 8th May 2020 for prediction. If the size of the training dataset is increased beyond 8th May 2020, there is a sudden drop of values in the recovered cases from the mid of May 2020 which indicates the error. The same kind of error was achieved in all the prediction models for the recovered cases as shown in Fig. 13. Country and city-specific analysis of COVID-19 prediction are done in the subsequent sections.

Country and City-specific predictive analysis – a case study

The previous Section discusses the predictions of COVID-19 cumulative confirmed, deaths and recovered cases for world-wide data. A comparative analysis of four different models are done and Stacked LSTM performs well than LSTM, ARIMA and PROPHET models. The results are arrived based on the data and the parameter settings of the model. Since the models predict cumulative cases accurately globalwide, models can be used to predict the country-wise and region-wise COVID-19 cases accurately. In this Section, the forecasting of COVID-19 cases is done as a case study in India (Country) and Chennai (City).

Predictive analysis for COVID-19 cases in India

As India is one of the countries that have higher infected cases, predicting the infection status would be pivotal for influencing key decisions. The prediction tools would help the public health administration in identifying the number of cases in the near future for taking preventive measures accordingly. The data from 22/1/2020 till 12/8/2020 has been considered from which 80% of the data is used for training and the remaining 20% is used as test data. The dataset of India was taken from John Hopkins University (JHU) [46] for analysis.

For this case study, we also performed a comparative analysis of four models whose result is presented in Table 5. It can be inferred that SLSTM outperforms other models.

The data is also tested with other models like LSTM, ARIMA and Prophet and MAPE values of ARIMA model is 2.08, 2.76 and 1.21 for confirmed, deaths and recovered cases, respectively. The MAPE values of LSTM model is 0.43, 1.9 and 2.4 and for Prophet model, the values are 2.8, 3.7, 2.7 for confirmed, deaths and recovered cases, respectively.

SLSTM model is tested on the dataset of India from 22nd January 2020 till 8th May 2020. The model is tested for daily predictions where the average percentage of error for the test data till 8th May 2020 is 0.02% and the actual and predicted values are shown in Fig. 14(a). Fig. 14(b) shows the future prediction of confirmed cases and at the end of August 2020, the cumulative confirmed cases in India is around 3.8 million. Test data was considered from 11/8/2020 till 20/8/2020.

Fig. 15(a) shows the daily predictions for the test data and the average error in percentage is 0.03% for death cases. The predicted value of cumulative death cases in India at the end of August 2020 will be around 69,477 cases which are plotted in Fig. 15(b). The average

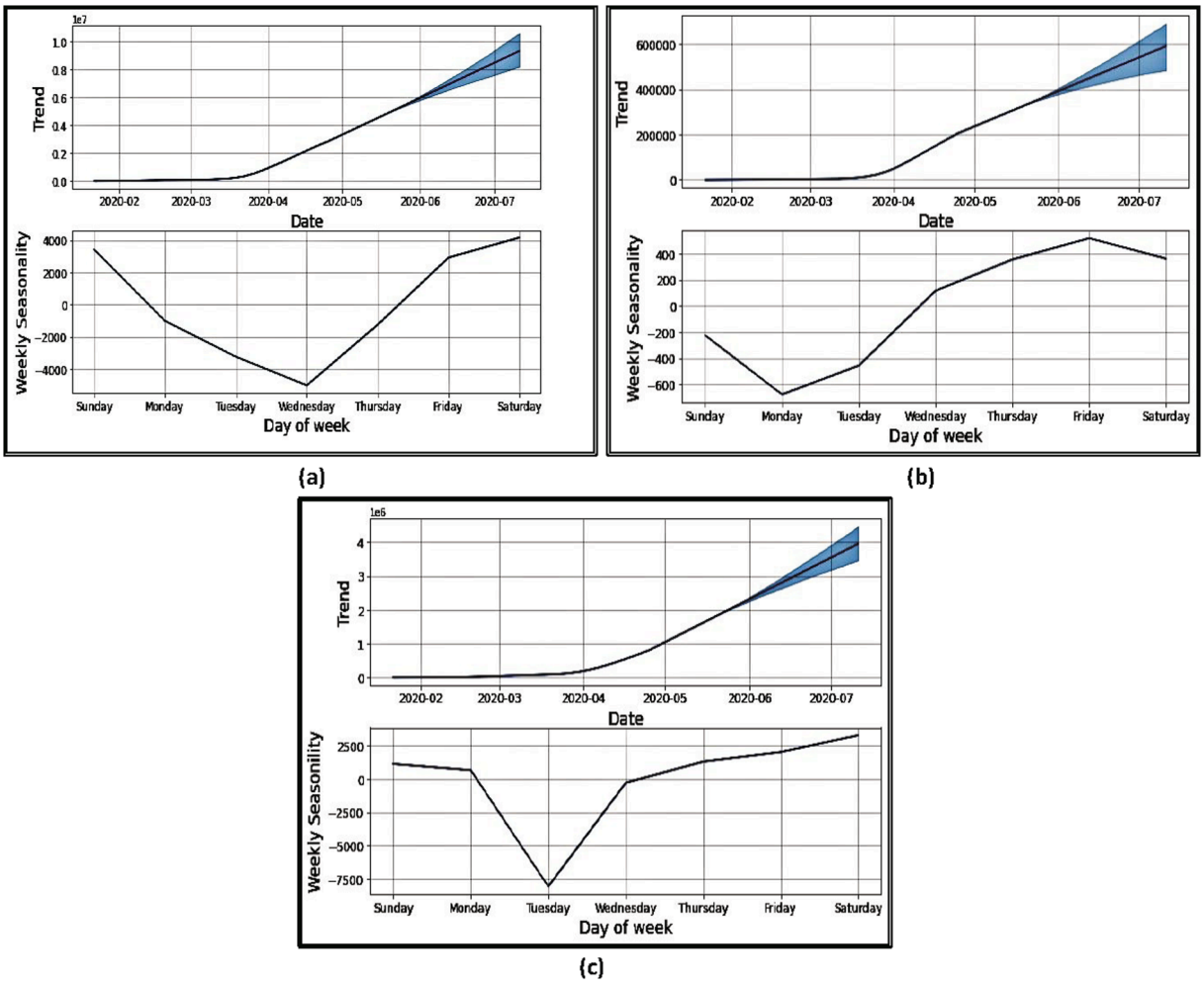

Fig. 9. Forecast components of Prophet Model in predicting (a) confirmed, (b) death and (c) recovered cases (until end of June 2020).

Table 3
Performance Evaluation.

Model	Predicted variable	RMSE	MAE	MAPE	R^2
ARIMA	Confirmed	10078.36	8097.55	0.372	0.94
	Deaths	1359.27	1067.46	0.742	0.938
	Recovered	8806.29	6551.99	1.12	0.92
LSTM	Confirmed	10051.22	9201.02	0.37	0.964
	Deaths	1670.84	1366.66	0.53	0.97
	Recovered	14210.39	12370.8	1.07	0.95
SLSTM	Confirmed	9310.83	7218.97	0.2	1
	Deaths	1219.35	1102.21	0.43	0.998
	Recovered	13201.4	11675.9	0.9	0.92
PROPHET	Confirmed	11516.2	8154.4	0.39	0.92
	Deaths	1348.71	1056.5	0.7	0.90
	Recovered	24485.6	17435.5	1.2	0.88

error in percentage for recovered cases is 0.02% and the predicted recovered cases at the various time period is shown in Fig. 16(a) and (b).

Predictive analysis for COVID-19 cases in Chennai (City)

Chennai is one among the most infected cities in India (present in the state of Tamil Nadu) and hence, carrying out the predictive analysis in this city would trigger the response in advance. The city reported its first COVID-19 case on March 18th, 2020 and recorded increasing clusters of cases in April and May. The government implemented rules and preventive measures like lockdown and social distancing to prevent the spread of COVID-19. On June 30th, there were about 2393 fresh cases and the total went up to 55,969 cases and 21,681 patients were recorded as active cases. On July 31st 2020, Chennai recorded 1013 active cases with a total of 98,767. In this case study, prediction of confirmed, recovered and death cases in Chennai is discussed. SLSTM model outperforms the other three models in the prediction of global analysis and country-specific analysis. Hence, only the results of stacked LSTM for Chennai COVID-19 prediction are considered.

Figs. 17–19 represents the future forecast values of confirmed, death and recovered cases using SLSTM model. The dataset for Chennai was generated from the data available at Indian COVID-19 website [53]. The cumulative dataset from 20th May 2020 till 20th August 2020 was considered for prediction. Starting from 11/8/2020 till 20/8/2020 was considered as test dataset. SLSTM model predicts the future cases accurately and the achieved MAPE values are 0.267%, 0.266% and 0.312% for confirmed, death and recovered cases, respectively. The highly accurate short-term predictions can help the public authorities to take decisions on lockdown measures and other economic activities.

Therefore, we can use the models for predicting global, country and city-specific infected cases and each prediction has its own significance. The global prediction provides quality information for humanity to assess their overall response to the pandemic. Meanwhile, the country and city-specific predictive analysis are much supportive in making decisions for economic recovery and detailed aspects of practical potentiality for prediction is briefed in the subsequent Section.

Deep learning models achieve good forecasting performance in

Table 4
Comparison of Actual and Forecasted cases.

Date	Observed Values	Forecast Values							
		ARIMA	Error (%)	LSTM	Error (%)	SLSTM	Error (%)	PROPHET	Error (%)
Comparison for confirmed cases									
5/27/2020	5,700,405	5,813,095	1.94	5,802,626	1.79	5,717,622	0.3	5,429,379	4.75
5/28/2020	5,819,719	5,907,110	1.48	5,929,106	1.88	5,835,205	0.27	5,513,519	5.26
5/29/2020	5,940,890	6,058,147	1.94	6,005,139	1.08	5,956,332	0.26	5,598,714	5.76
5/30/2020	6,078,719	6,113,348	0.57	6,189,701	1.83	6,081,633	0.05	5,677,062	6.61
5/31/2020	6,186,277	6,324,107	2.18	6,218,889	0.53	6,211,114	0.4	5,756,597	6.93
Comparison for death cases									
5/27/2020	359,038	383,378	6.78	374,025	4.17	352,072	1.94	386,796	4.75
5/28/2020	363,749	387,903	6.64	379,073	4.2	356,317	0.04	392,737	5.26
5/29/2020	368,496	393,712	6.84	383,378	4.04	360,513	2.17	398,760	5.76
5/30/2020	372,662	400,116	7.37	387,903	4.09	364,671	2.14	404,615	6.61
5/31/2020	375,555	405,840	8	393,712	4.83	368,793	1.8	409,812	6.93
Comparison for recovered cases									
5/27/2020	2,346,232	2,263,404	3.53	2,423,004	3.27	2,334,784	0.49	1,976,908	15.7
5/28/2020	2,413,089	2,316,140	4	2,537,655	5.16	2,387,862	0.05	2,015,088	16.4
5/29/2020	2,490,435	2,367,520	4.94	2,658,750	6.73	2,441,177	1.97	2,051,215	17.6
5/30/2020	2,560,888	2,431,633	5.05	2,786,033	8.79	2,494,732	2.6	2,084,831	18.5
5/31/2020	2,637,208	2,485,323	5.8	2,916,419	0.5	2,548,528	3.36	2,118,968	19.6

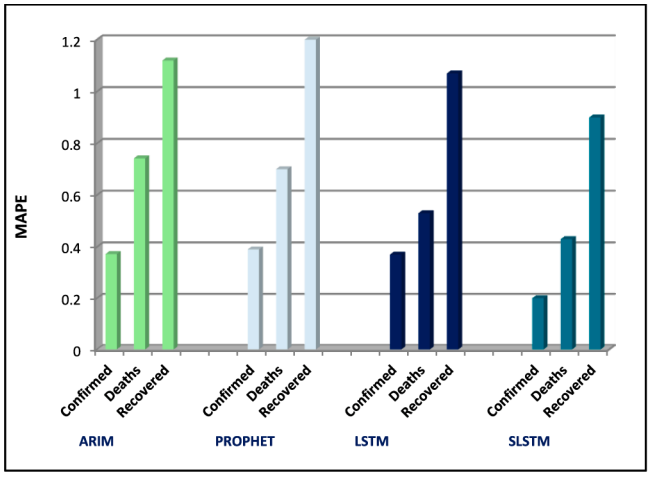

Fig. 10. MAPE Comparison of four models.

handling the time-series dataset. Deep learning models are demonstrated for the prediction of COVID-19 cases. Based on the graphical results and the performance metrics, SLSTM is better than the other models in forecasting the pandemic infection status world-wide. The input data is made suitable to the model and two-layer stacked LSTM with 100 neurons is used along with the return sequences set as true. For connecting each neuron with the next neuron in a fully connected network, one dense layer is added. The bias due to random initialization is reduced and based on the best number of hidden layers and hidden states were chosen, SLSTM performed better compared to other models with lower RMSE and MAPE values. LSTM outperforms ARIMA and Prophet in predicting confirmed and death cases. The results are purely based on the data and the parameter settings of the model, there is no obvious answer that one of the algorithms is always better than the other in all cases. However, for this application SLSTM outperforms other models in forecasting COVID-19 prediction. Since other models are also reasonably good, statistical analysis and hypothesis testing is done to select the best suitable model.

Statistical analysis

A statistical hypothesis test called T-test is carried out to find out the best suitable prediction model. T-test hypothesis used in this study is the ratio of the difference between the two means and the measure of variability or dispersion of groups. The t-value is calculated using Eq. (23),

$$t = \frac{\bar{X}_T - \bar{X}_c}{SE\left(\bar{X}_T - \bar{X}_c\right)} \tag{23}$$

The numerator indicates the difference between the mean value and the denominator indicates the standard error. If the computed value is below the threshold value of statistical significance ($\alpha = 0.05$), then the null hypothesis is rejected and it is accepted if the threshold is greater than the α value. The performance of the prediction models used in this study was investigated based on the dataset like confirmed, death and recovered cases taken from John Hopkins University. Hypothesis testing is performed on the time series datasets to check if the data is stationary or not. A comparative analysis of the designed models is performed based on the results of hypothesis testing. The models can be used to predict COVID-19 cases of any country [32]. The results are shown in Table 6 and the best performing SLSTM model is compared with Prophet, LSTM and ARIMA models respectively and the test results are described in the table.

From the statistical analysis, it is concluded that SLSTM, ARIMA and LSTM are acceptable for forecasting COVID-19 cases. SLSTM outperforms other models. Also, the ranking performance has been analyzed for each of the models based on CC, MAE and RMSE. The best performing model is assigned the rank of 1 and the worst is ranked 0. The rank on the j^{th} algorithm on the i^{th} dataset is computed according to Eqs. (24) and (25).

$$R_{ij} = 1 - \frac{e_{ij} - \min(e_i)}{\max(e_i) - \min(e_i)} [\text{For MAE and RMSE, lower values are better}] \tag{24}$$

$$R_{ij} = 1 - \frac{e_{ij} - \max(e_i)}{\min(e_i) - \max(e_i)} [\text{For CC, higher values are better}] \tag{25}$$

where e_{ij} is the measured value for the jth algorithm on dataset i, and e_i is the vector accuracy for dataset i. The correlation coefficients are computed for the models for confirmed, death and recovered cases [54].

In terms of the correlation coefficient, SLSTM with the value (1.0) performs well than ARIMA (0.9) and LSTM (0.5) for confirmed cases, and similarly, for the death cases, LSTM (0.9) performs well than the other models and SLSTM (1.0) performs well for the recovered cases. From the above analysis, it can be inferred that from the preliminary modelling performance, no model performs well for all the measures or

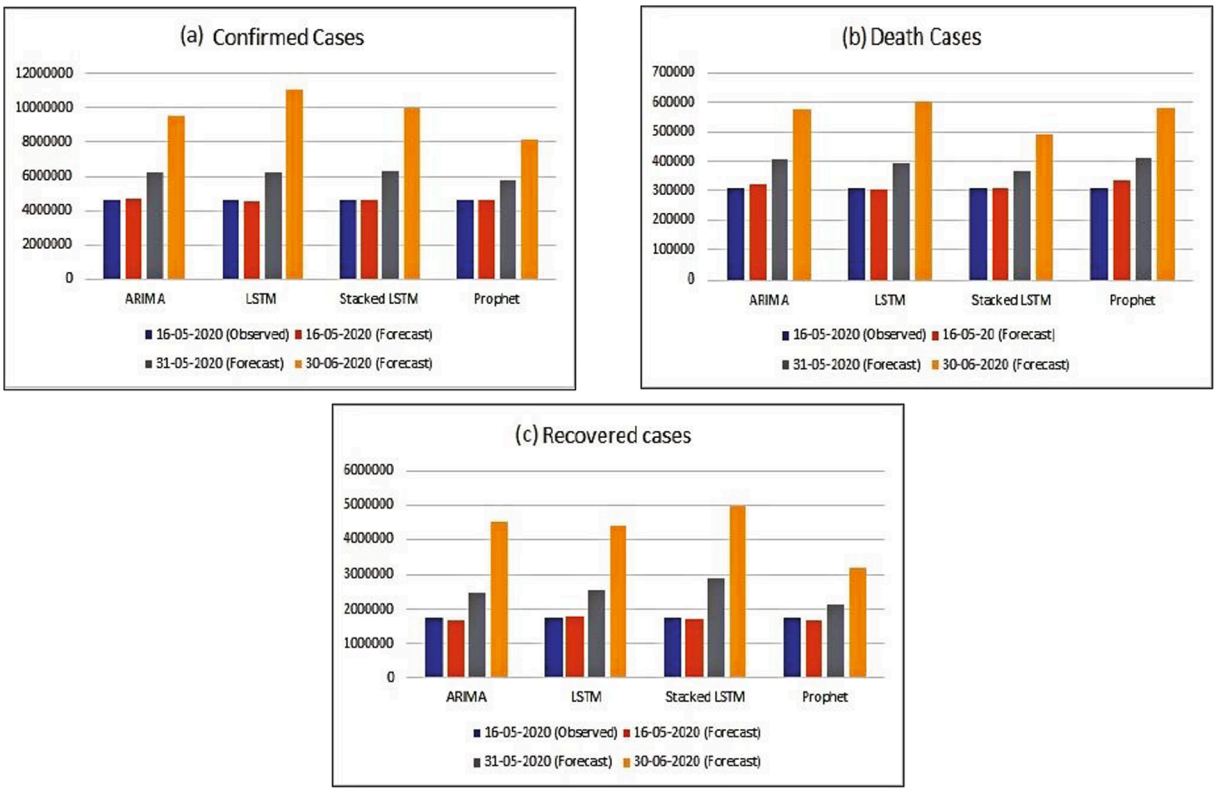

Fig. 11. Model-wise comparison of (a) confirmed, (b) death and (c) recovered cases.

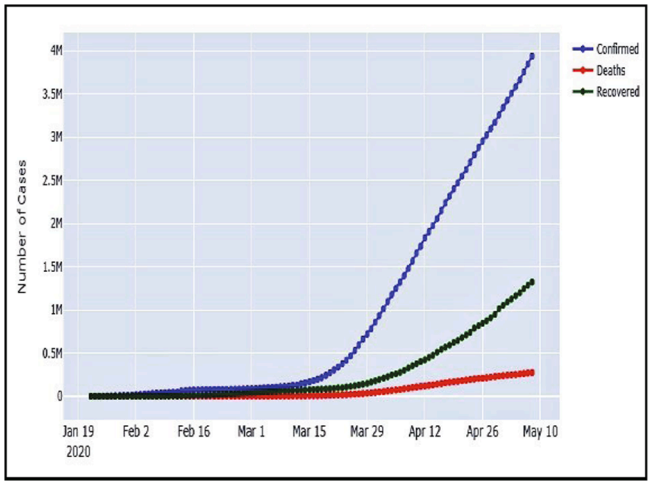

Fig. 12. Predicted COVID-19 cases.

attributes. So, it is difficult to select the best model from the above analysis. According to the results, no single model is 100% accurate in all aspects of prediction. Therefore, ranking performance can be estimated using the above equations to select the best performing model. The average ranking of the performance metrics for different models is computed using the Eqs. (24) and (25) and ranks are represented in Table 7 and Fig. 20.

Overall, as demonstrated by the RMSE and MAE measured performance and the statistical ranking, SLSTM outperforms other models because of the best hyperparameter tuning and reduction in bias and ARIMA outperforms LSTM model. Here, the general trend of the data is considered and the prediction of COVID-19 cases helps us to be aware of the future cases and to take necessary actions to alleviate it.

Correlation analysis for COVID-19 prediction

The relationship between variables can be captured using correlation analysis. Correlation may be positive, negative or neutral. A positive correlation indicates the movement of variables in the same direction. A negative correlation indicates the change of variables in the opposite direction and neutral correlation indicates no relationship in the change of variables. Different correlation scores can be obtained based on the distribution of variables. The linear relationship between variables can be captured using the Pearson correlation. The Pearson's correlation coefficient is calculated as the covariance of the two variables divided by the product of the standard deviation of each data sample. It is the normalization of the covariance between the two variables to give an interpretable score. Correlation factor "r" can be identified by the color variation and the change of color indicates the weak correlation to the strong correlation. The value of r ranges from -1 to $+1$ as shown on the side of the correlation matrix.

In this section, the correlation analysis is performed to capture the relationship between the COVID-19 cases and other external factors such as temperature, rainfall, population etc. If the value of the correlation $|r| > 0.6$, it indicates a strong correlation. The features considered for correlation are total population, area, population density, COVID-19 cases, temperature and rainfall during the months of May, June, July, August, and September 2020 for seven cities in Tamil Nadu. The data is obtained from [55]. The feature correlation is identified using the correlation heat map. The correlation analysis illustrated in Fig. 21 represents that the COVID-19 cases have the dependency on the dynamic features like temperature and population. There is a strong correlation between temperature and COVID-19 cases.

Also, the correlation analysis for Chennai COVID-19 cases and daily average temperatures are explored using Pearson's correlation as represented in Eq. (26)

$$\mathrm{Cov}(X,Y)/(\mathrm{std}(X) * \mathrm{std}(Y)) \tag{26}$$

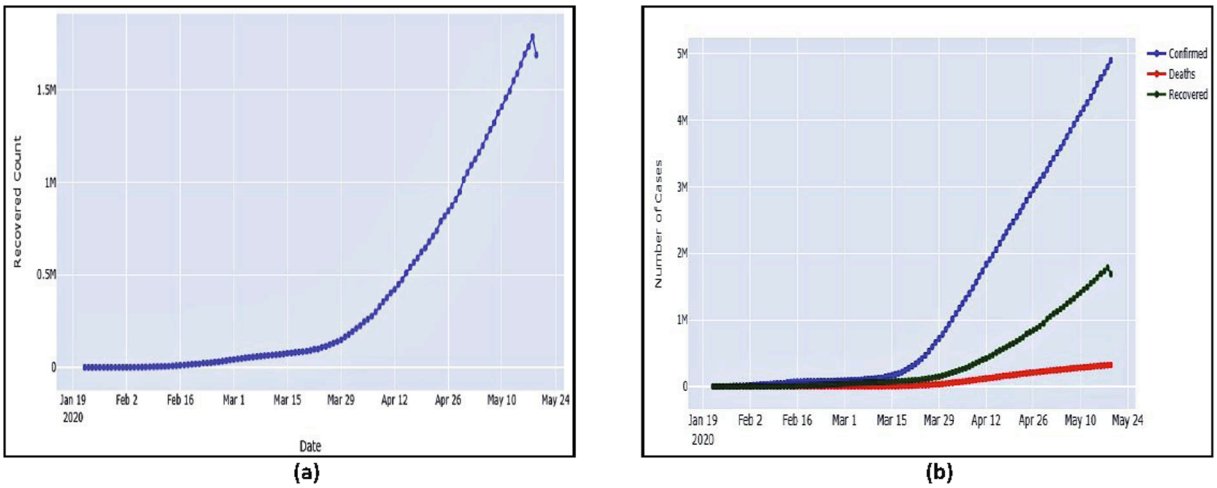

Fig. 13. Plot of error (a) in recovered cases (b) overall prediction.

Table 5
Performance Evaluation for country-specific prediction.

Models	Predicted variables	RMSE	MAE	MAPE
ARIMA	Confirmed	194.76	70.94	2.08
	Deaths	2324.9	1350.86	2.76
	Recovered	2178.86	1305	1.21
LSTM	Confirmed	1167.56	979.03	0.43
	Deaths	992.84	866.66	1.9
	Recovered	1670.84	1366.66	2.4
SLSTM	Confirmed	274.22	920.02	0.3
	Deaths	309.12	278.29	0.6
	Recovered	1125.47	864	1.8
PROPHET	Confirmed	9970.33	7231.1	2.8
	Deaths	1843.71	1389.5	3.7
	Recovered	9310.83	7633.49	2.7

and Spearman's correlation is computed based on the rank of values of the samples and is denoted in Eq. (27)

$$\mathrm{Cov}(\mathrm{rank}(X),\mathrm{rank}(Y))/(\mathrm{stdv}(\mathrm{rank}(X)) \quad (27)$$

It is clear from Fig. 22(a)–(c), there is no strong correlation between daily average temperature and COVID-19 cases in Chennai. The dataset from May 2020 till August 2020 is taken for analysis and the Pearson, Spearman and Kendall correlation coefficient values are 0.41, 0.25 and 0.2, respectively. The values show that there is a weak correlation between COVID-19 cases and other external factors. From the above study, it can be inferred that there may be a direct influence excerted by important factors in terms of climatic and geographic characteristics or may not have an impact on COVID-19 cases. The analysis of seven different cities in Tamil Nadu shows that the weather parameters affect the COVID-19 cases and the analysis of Chennai with daily average temperature shows a weak correlation with the number of COVID-19 cases. Also, the population and the confirmed cases have a positive correlation and may have a high chance of getting a greater number of cases. As the coronavirus is highly transmittable, people should understand the importance of lockdown, social distancing and social isolation to prevent the spreading of the pandemic.

Multivariate stacked LSTM model for COVID-19 prediction

In this section, a case study of multivariate LSTM model is demonstrated by considering the combined dataset with multiple variables like confirmed cases, death cases, recovered cases, latitude and longitude. Stacked LSTM is used to predict the increasing rate of COVID-19. Different regions (countries/provinces/states) are used as the training data and one country as the validation data. Dataset is taken from John Hopkins University from 22nd Jan 2020 till 11th Nov 2020 and dataset is divided into training and testing data. Min Max scaling is done to normalize the data and the data is transformed to the same scale to

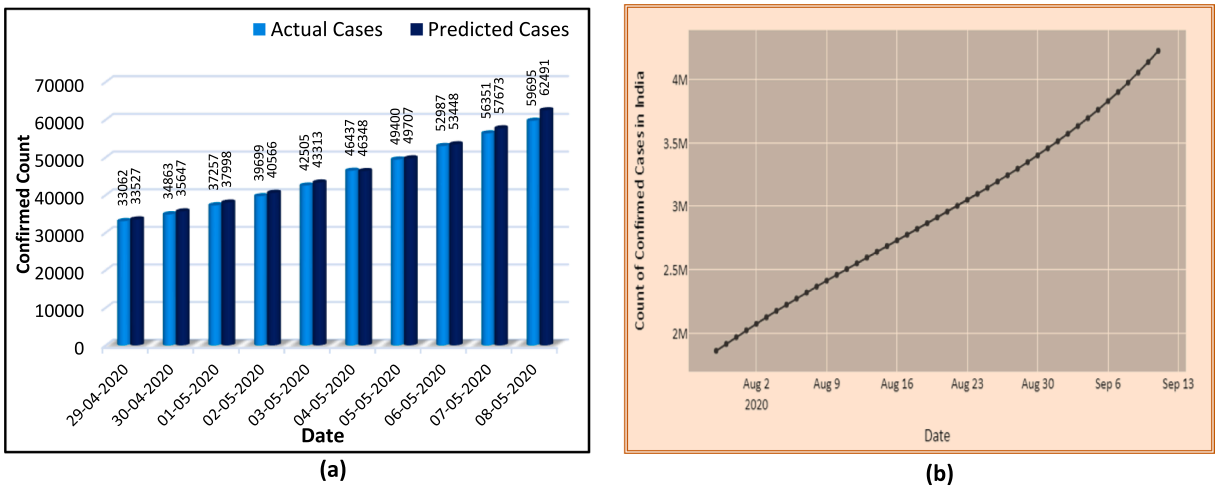

Fig. 14. Prediction of Confirmed cases in India (a) Monthly comparison (b) Forecasted data upto August 2020 (Dataset was taken till 20/8/2020).

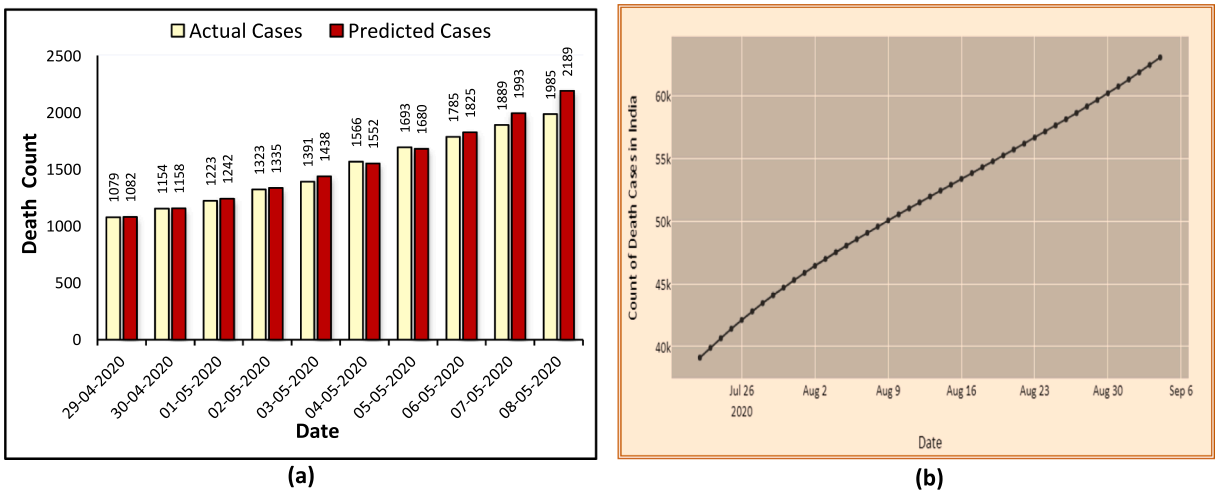

Fig. 15. Prediction of Death cases in India (a) Monthly comparison (b) Forecasted data upto August 2020 (Dataset was taken till 20/8/2020).

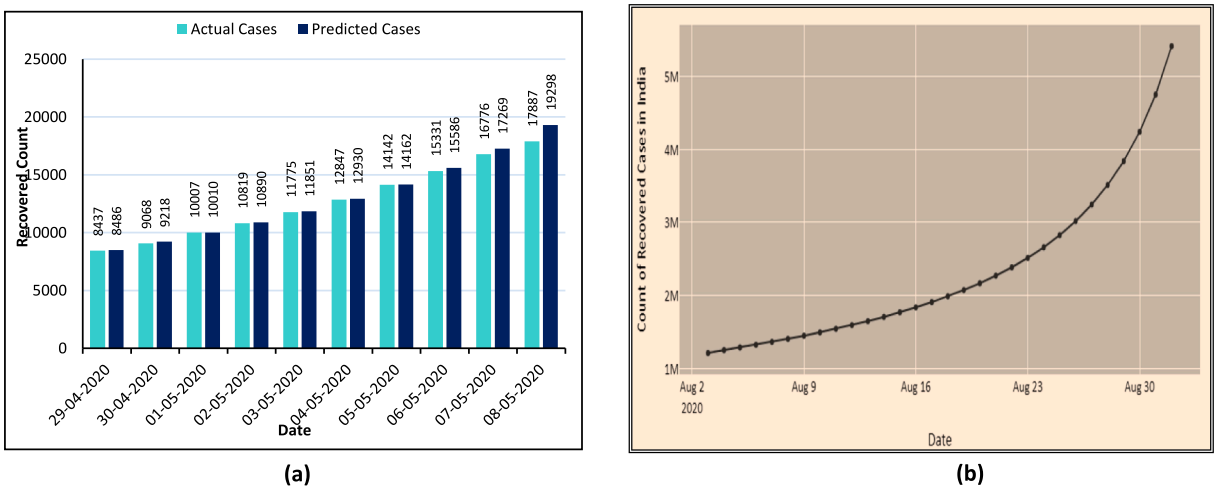

Fig. 16. Prediction of Recovered cases in India (a) Monthly comparison (b) Forecasted data upto August 2020 (Dataset was taken till 20/8/2020).

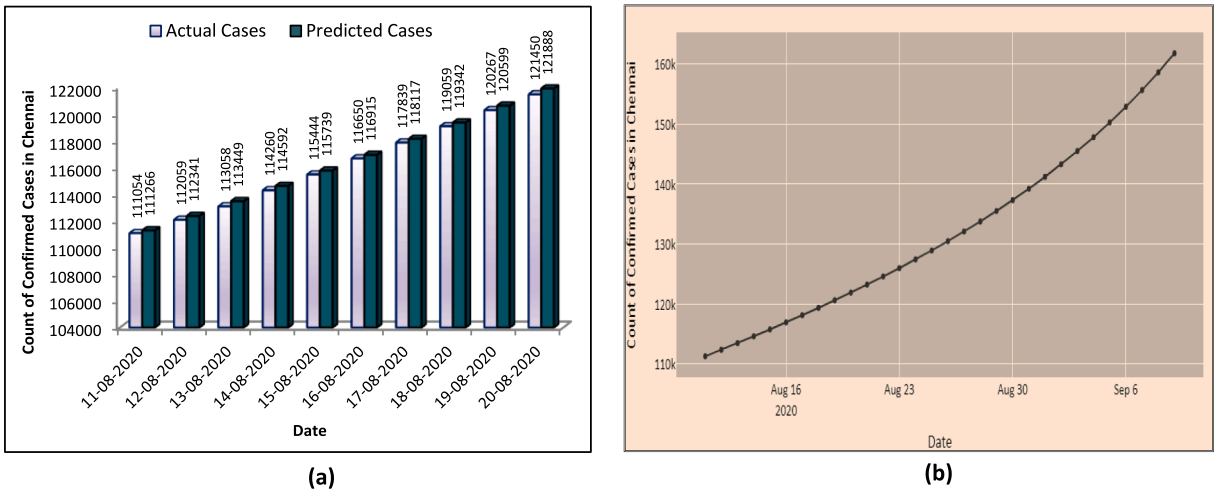

Fig. 17. Prediction of Confirmed Cases using SLSTM in Chennai (a) Monthly Comparison (b) Forecasted upto mid-September 2020 (Dataset was taken from 20/05/2020 till 20/08/2020).

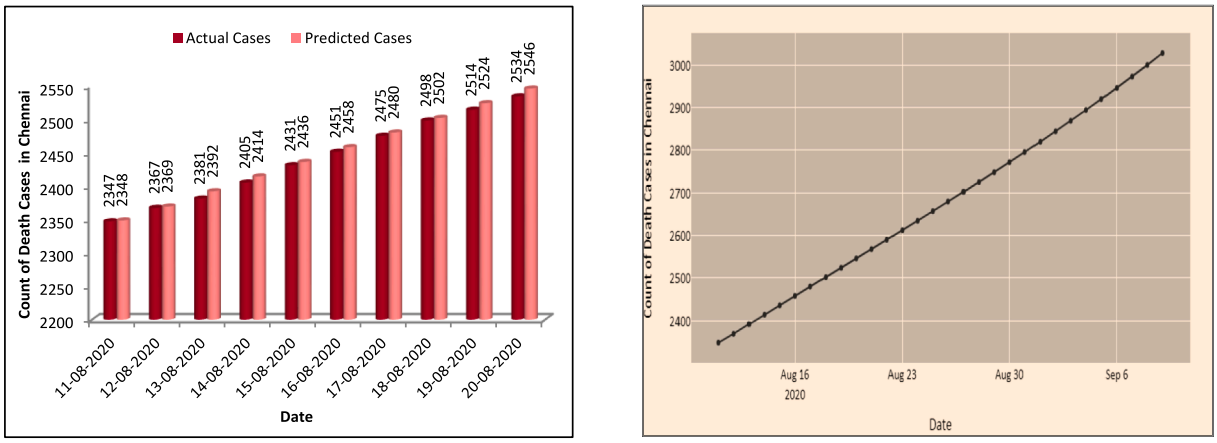

Fig. 18. Prediction of Death Cases using SLSTM in Chennai (a) Monthly Comparison (b) Forecasted upto mid-September 2020 (Dataset taken from 20/05/2020 till 20/08/2020).

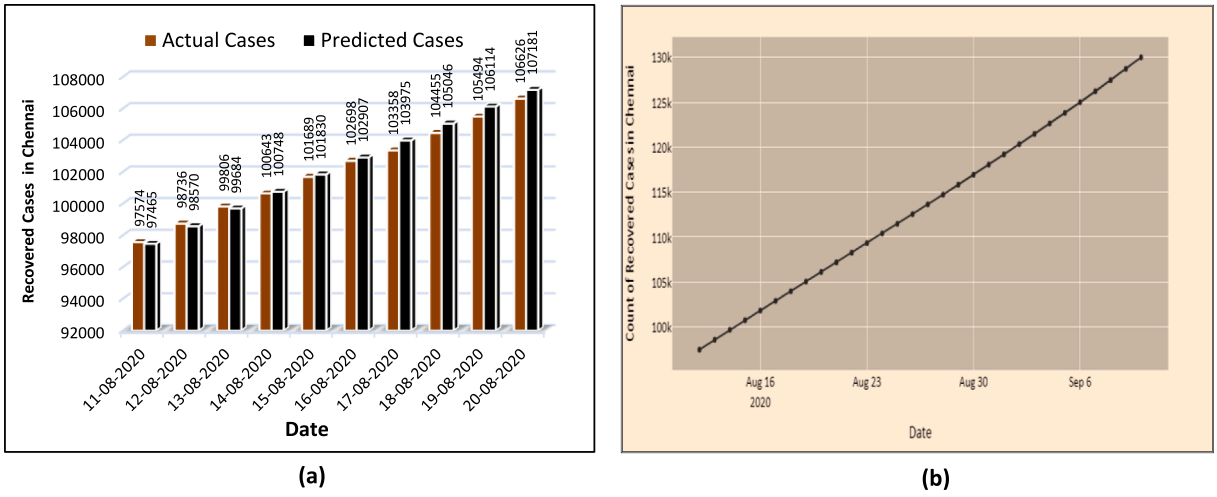

Fig. 19. Prediction of Recovered Cases using SLSTM in Chennai (a) Monthly Comparison (b) Forecasted upto mid-September 2020 (Dataset was taken from 20/05/2020 till 20/08/2020).

Table 6
Results of hypothesis testing.

Time series data	Test Result	Hypothesis status
Total confirmed cases	Statistics: 0.512 p-value: 0.679	Fail to reject H0
	Statistics: 0.306 p-value: 0.763	Fail to reject H0
	Statistics: 0.058 p-value: 0.954	Fail to reject H0
Total recovered cases	Statistics: −38.457 p-value:0.000	Reject H0
	Statistics: 0.435 p-value:0.678	Fail to reject H0
	Statistics: 0.290 p-value:0.775	Fail to reject H0
Total death cases	Statistics: −37.475 p-value: 0.000	Reject H0
	Statistics: 0.431 p-value:0.671	Fail to reject H0
	Statistics: 0.374 p-value:0.713	Fail to reject H0

Table 7
Ranking of Algorithms,

Model	Average Ranking	Overall Rank
ARIMA	1.9100	2
LSTM	2.1134	3
SLSTM	1.7531	1

reduce the bias. Adam optimizer with the learning rate of 0.0001 with the value of training episodes 10,000 is used. Also, the bias is reduced by training and testing multiple times that occurred due to random initialization. Fig. 23 shows the validation result of India. The number of confirmed cases increases and the growth pattern is depicted in the Fig. 23. The optimal parameters are chosen and the hidden layers are increased heuristically to see the impact of prediction accuracy. The number of confirmed cases at the end of October 2020 is 8.1 million and will cross 9.2 million at the end of November. The number of recovered cases is 7.4 million and 9 million at the end of October and November respectively.

Based on the parameters used in this study and considering the different architectures, RMSE values decreases if the number of hidden state increases. The model is trained to use the world-wide data

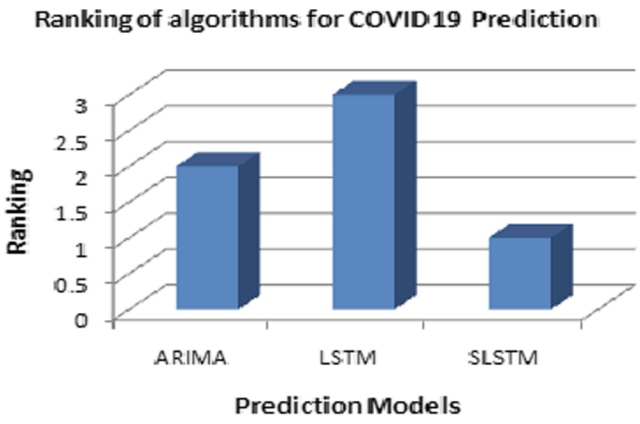

Fig. 20. Ranking of algorithms.

optimally. RMSE value initially obtained is 2712.1 for the number of hidden states as 1, 865.2 for 5 hidden layers, 673.6 for 10 hidden layers and 486.1 for 30 hidden layers. The performance of the stacked LSTM model increases as the number of layers is increased. The total prediction sequence is 100 days and 92% prediction accuracy is achieved using multivariate analysis.

Therefore, we can use the models for predicting global, country and city-specific infected cases and each prediction has its own significance. The global prediction provides quality information for the humanity to assess their overall response to the pandemic.

The forecasting models with higher accuracy will be very much helpful for the health care system. There are many challenges in forecasting COVID-19 since the longer incubation period with very few available datasets are present. If the model can be utilized to forecast accurately, then it will be helpful for the decision makers to plan about proper lockdown, lockdown period, can educate people to be aware without panic, to maintain social distancing and making the essential services available before lockdown. If the model does not perform well, it may affect the overall performance and mislead the prediction.

Meanwhile, the country and city-specific predictive analysis is much supportive in making decisions for economic recovery and detailed aspects of practical potentiality for prediction is briefed in the subsequent Section.

Unleashing the practical potentiality of prediction during COVID scenario

The main objective of the study is to compare the four models and derive the best out of it. But there exists a question for the purpose of such predictions, especially during COVID scenario. Thus, this section is dedicated for assessing the practical significance of prediction of infected cases and also, the potentiality of prediction in the society.

Fig. 21. Correlation between COVID cases and external factors.

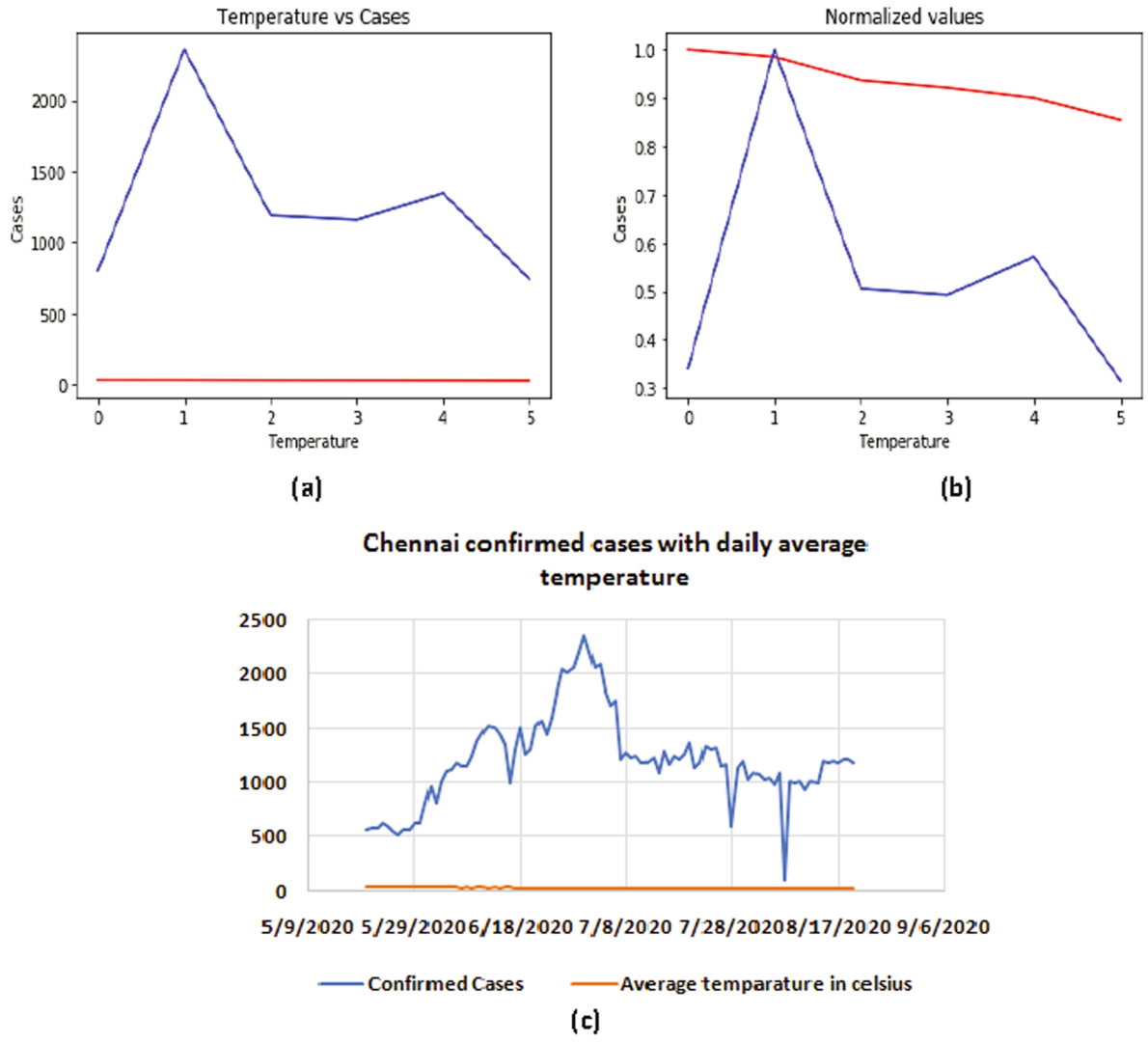

Fig. 22. (a) Chennai COVID 19 cases vs Temperature, (b) Normalized values for temperature and cases and (c) Chennai Confirmed cases vs daily average temperature.

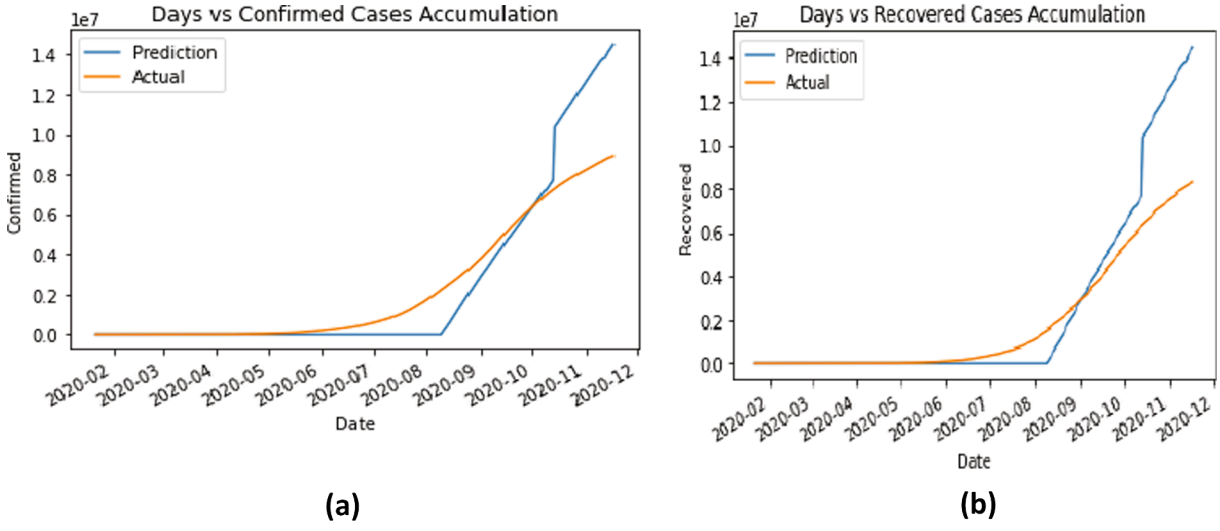

Fig. 23. Confirmed and recovered cases of India.

Assessing the pandemic characteristics in the predicted region

The data which the authors are predicting include confirmed cases, death and recovered cases. With all these data, active cases can be determined with the Eq. (28).

$$\text{Active cases} = \text{Confirmed cases} - (\text{Death cases} + \text{Recovered cases}) \quad (28)$$

The rate of change of active cases is very essential to determine the intensity of infection scenario occurring in society. It effectively maps the time when it would reach the peak and thus, influences the governmental measures. On the other hand, the recovery characteristics of certain locality depend on the population immune nature to the disease, the virulent strain of the pathogen, and especially the age group. Characterizing the age group presented within the locality would ultimately help in predicting the mild and critical cases that would probably emerge among the infected cases [56,57].

Assessing the pandemic impacts in medium and long-term is necessary. For such a case, the duration of the pandemic in a region needs to be predicted so that the measures can be planned accordingly. Hence, the growth rate act as a crucial parameter to assess the duration of pandemic occurrence. If it is positive, the slope of the epidemiological curve of infection is most likely to increase while if it is negative, the vice-versa would occur. On the other hand, the peak infection duration is especially needed for governments to start implementing recovery measures. These characteristics of a pandemic can be well-determined with the predicted data. However, the overall duration of pandemic occurrence and peak infection duration data is less accurate to predict at an early stage of infection spread.

The huge differences in confirmed cases between predicted and actual value most probably would result from some virus hotspot that needs to be identified as this is highly helpful to trace the individuals else, community spread would emerge. On the other hand, if there is a drastic difference in the death rate, there is a possibility that the virus has mutated into much virulent strain. This would even trigger the research community to assess whether a new strain of coronavirus has evolved as this is significant for the preparation of the vaccine. Meanwhile, if the recovery rate experiences an upheaval of difference, then either the immunity of population would have got better or the government measures excelled or the virus virulence would have decreased suddenly. From all these, it can be inferred that the predicted data can act as a reference for a normal situation and any abnormal variations should be traced for its reason which would help the humanity in plentiful of ways.

Scenario planning

The prediction of COVID-19 cases is highly useful when the prediction is brought down to the city's scale. Various scenario planning within the society can be accomplished if the infected cases are predicted earlier.

Prediction of possible active cases ahead of time would help to shape the society's response in advance to minimize the impacts. The active cases would help the healthcare sector a lot. For instance, the society's or city's healthcare capacity should be always less than the active cases, especially the symptomatic cases for effective treatment or else, the hospitals would flourish with patients and eventually results in massive deaths as occurred in Italy and U.S. Typically, mild infected cases would take 2 weeks to recover while critical cases would take 3–6 weeks. With this data, the occupancy of hospital beds can be effectively managed and prioritized for critical patients if the recovery characteristics of the population in the locality is known. Thus, if the active cases are determined in advance, then depending on the healthcare capacity, either the expansion of healthcare facilities can be done or effective management of patients can be encountered with the available healthcare system facilities. Further, imparting the population response in the locality, the recovery rate can be enhanced by giving prioritization to probable critical patients (diabetics, old age, and respiratory problems). The ventilators are crucial in treatment for symptomatic patients and ordering the required quantity in advance remains uncertain. This major issue can be solved with the prediction of active cases ahead of the month as the orders can be met as soon as possible. Even if the supply chain is ruptured, the orders can be planned accordingly, as predicting one or two months of data ahead would give us a good preparation period. Moreover, the testing kits if needed, can be imported from other countries by placing advance order and also, partnerships can be initiated to develop technology for effective tracing, given the span of prediction.

Active cases data would also influence the approach by the healthcare sector in terms of testing and technological influence in context with tracing. More the active cases, the higher is the probability that the growth rate of infection in the locality increases. Thus, testing and tracing methods can be implemented at leaping acceleration according to the prediction data. If a city tends to exhibit higher active cases, testing facilities can be expanded and advance measures to make it accessible to a large population can be explored. Regarding tracing, adopting a technologically oriented approach is promising when the active case rises drastically. The region or locality can be forced to use governmental tracing apps and many such measures would ultimately slow down the infection spreading rate. Thus, here the prediction plays a significant role to trigger the reaction in advance before it worsens.

From the government perspective, the prediction would influence the measures that are needed to be implemented. If the authorities in power discuss with experts regarding suggestions, then the decision is most likely would be based on evidence which is backed up by prediction data. The necessary measures include lockdown measures, supporting healthcare needs, planning stimulus packages, expanding healthcare facilities and framing strategies to mitigate the infection. For all this, predicted data would act as a reference and thus, measures can be implemented with ease so that the actual scenario rewards us with minimal impacts. On the other hand, risk communication can be effectively carried out supported by predicted data. Even awareness can be imparted with predicted data to get people's cooperation through various medium to adopt social distancing practices and to wear a mask as a preventive measure.

Governmental orders especially lockdown measures severely affect the society's normal functions. Industrial sector and employment get affected the most if the proper implementation of lockdown isn't accomplished. For instance, industries can manage with their inventories if they receive information of lockdown in prior while the companies and commercial services would take suitable actions to promote work from home strategies to ensure the continuity. Similarly, retail markets and people can prepare themselves to experience the lockdown scenario with a reasonable period of time as the measures can be announced earlier. This also helps in the smooth distribution of prior resources when compared to the sudden implementation of lockdown. As a summary, the society can be restructured smoothly without sudden changes if predicted data is interpreted seriously.

The energy sector is impacted especially in terms of demand variations owing to the lockdown. The residential load is increased as people stay at home while the industrial, commercial and transportation load decreases [58]. On the whole, the load demand is significantly reduced. But when the predicted data also influences the governmental measures, the regions of lockdown is well-known well ahead. This ultimately helps the energy sector to analyze the possible demand variations in a particular locality and act accordingly by implementing appropriate mitigation measures. Further, staff allocation and resource planning can be effectively accomplished.

Scaling up COVID-19 prediction to multiple cities in a state would offer us huge benefits. Individual city's analysis in a group of cities would produce varying results which eventually can help in the implementation of decentralized measures to effectively control the pandemic

as well as to maximize the freedom to the public. For instance, if City A is predicted to have a high level of active cases after a span of 2 weeks, meanwhile the neighboring City B and C see a drop in the number of active cases during the same span of time, then the City A resident's mobility can be restricted within the city so that the infection spreading to the other cities can be avoided. This type of approach can be very effective during the recovery phase (when the slope of the epidemiological curve decreases consistently) of infection where 'divide and conquer' approach should be followed by the government to implement different measures according to the infection status of the locality. Moreover, the integration of prediction of multiple city's infection data and combined Multi-Criteria Decision Analysis would help the government to take qualitative decisions with the priorities concerned to both health and economy. Further looking forward in a futuristic society, even the impacts on the economy, emissions, health and the like can be accurately predicted in line with the predicted cases, followed by the governmental orders that influence the changes in the society. Such a series of predictions become a multidimensional analysis with considerable social factors leading us to make a better decision and implement advance measures which all favors to the development of sustainable societies.

In the current scenario, though the prediction facilities exist and forecasting occurs, emphasizing their results and effective interpretation of the same is lacking. For example, BlueDot predicted the outbreak of the disease at Wuhan much earlier and warned [59]. If we had reacted to the warning at the outbreak region, possibilities for the disease turning into a global pandemic would be much lesser. Thus, it is clear that we deemphasize the technological advancement in scenario planning and utilizing the full potential of technology, especially the prediction would reward us a better life.

Optimization of models

COVID like situation generates a lot of data regarding the infection. Hence, this can be used as an opportunity to train new models to improve the accuracy of the model. Apart from it, different social factors can be used as inputs and correlation can be done with the disease and social factors. Such analysis was contributed in some studies that focus on various geographical factors, climate and population density with its influence on infection spreading nature [60,61]. Thus, many such correlations would help in providing different aspects to map the disease transmission dynamics and its characteristics with respect to various environmental and social factors and also, to tune the response accordingly.

Sustainable development goals (SDGs)

Predicting the infected cases would directly or indirectly help in the progress towards SDG 3 (Good health and well-being) and SDG 11 (Sustainable cities). With the consideration of predicting active cases weeks or months ahead, the healthcare sector can prepare themselves to manage the upcoming scenario and outperform well to provide quality treatment. Thus, prediction offers the most valuable time for the healthcare sector to react correctly and this contributes towards SDG 3.

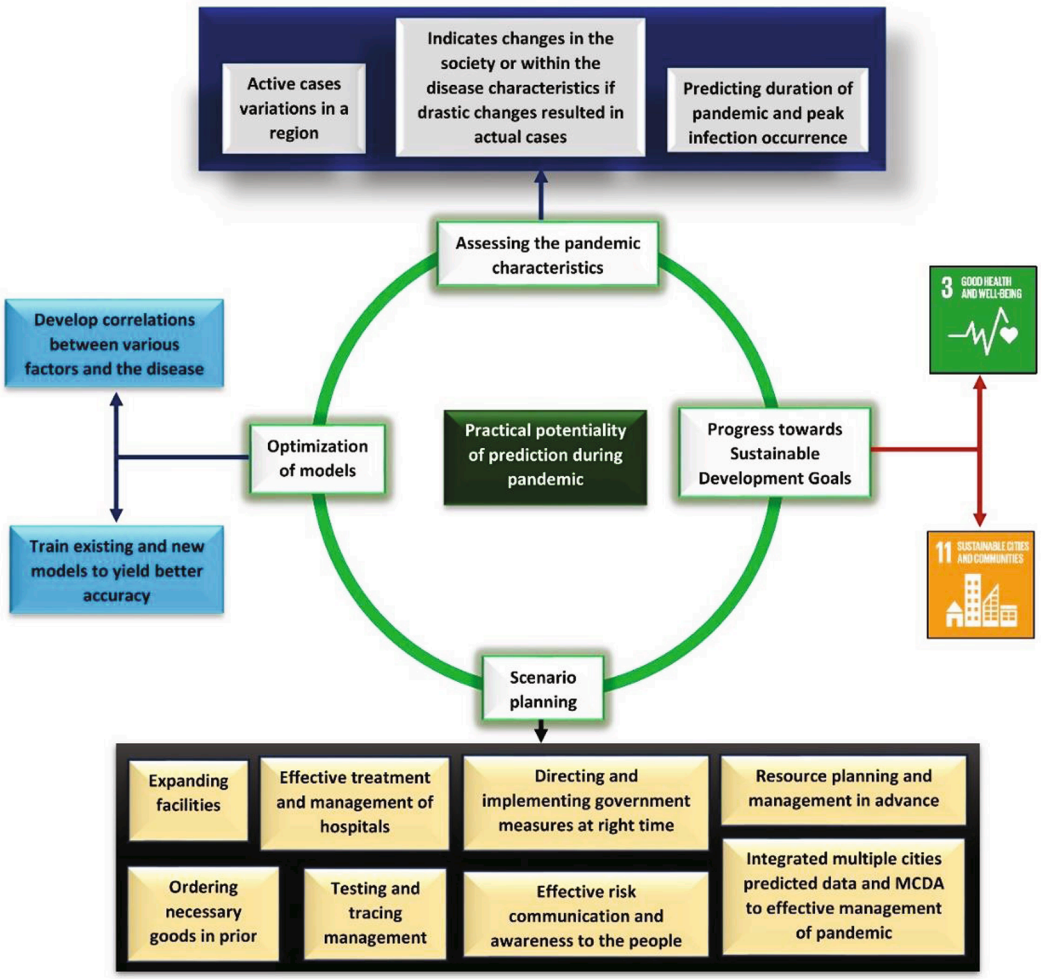

Fig. 24. Representation of practical significance for predicting COVID-19 infected cases.

The predicted confirmed data also influences the governmental actions which are followed by societal changes such as a lockdown. Implementing the right measures at the correct time would save millions of lives and thus, the prediction empowers the decision to be taken at the correct time. On the other hand, authors argue that from the societal influence of the predicted data stated above, one could say that the pandemic can be effectively handled with best decisions and minimal impacts on humanity. This puts us right in the pathway of sustainable cities (SDG 11). Prediction model has a direct impact on SDG 3 and SDG 11 but it also indirectly helps in reducing the negative challenges on other SDGs due to COVID such as hunger/poverty, economic disaster, energy and the like, if we can handle the COVID influences properly.

Fig. 24 represents the practical potentiality of the prediction during COVID scenario. From the practical significance, it can be inferred that predicting data for long-term (months) would be helpful for preparedness and ordering of essentials while predicting data for short-term (weeks and days) would be supportive in decision making, implementing measures and changes to be occurred in the society to tackle the pandemic.

Furthermore, global cases prediction is useful in the statistical analysis of economic factors that requisites the infection status and to check the reliability of proposed mathematical models. The country-specific predictive analysis would influence international supply chain-related decision and trading, intense of economic recession, centralized measures to be adopted, testing and tracing methods. City-based predictive analysis yields diverse application and control over the society and moreover, accurate interaction within multiple cities would yield a better response to the pandemic.

Conclusions

The COVID-19 pandemic is a big threat to humanity and its damage to society is irreplaceable. Research is being carried out to minimize the loss of human lives by predicting the spread of pandemic outbreaks and in identifying vaccines. Accurate prediction of COVID-19 using deep learning has gained more attention in the current scenario. Deep learning methods are more significant in handling non-linear problems effectively. In this work, time series prediction of COVID-19 outcomes is done using ARIMA, LSTM, SLSTM and PROPHET models to estimate the future prediction of confirmed, death and recovered cases for the specified time intervals provided in the model. The proposed methodology is used for predicting both short-term and medium-term infected cases. The results of the analysis show that the Stacked LSTM and LSTM models outperformed other studied models with higher accuracy and it proves the reliability for predicting COVID-19 cases. During the fast spread of the pandemic, the Stacked LSTM models shows accurate prediction with MAPE values of 0.2, 0.43 and 0.9 for confirmed, death and recovered cases, respectively for global data analysis. The following are some of the conclusions extracted from the global data analysis.

- The expected total number of confirmed cases was predicted to be around 6.3 million at the end of May 2020 and around 9.9 million at the end of June 2020 which was very close to the actual value.
- The expected recovery cases and death cases was forecasted around 5 million and 492,261, respectively at the end of June 2020 which also resembled the actual value.

Predictive analysis for country and city-specific case study for India and Chennai was analyzed and the following prediction results are extracted from the analysis.

- The expected total number of confirmed, death and recovered cases in India will be around 4.3 million, 60,226 and 3.9 million at the end of August 2020.
- The expected total number of confirmed, death and recovered cases in Chennai will be around 137,309, 2,771 and 116,932 at the end of August 2020.

Statistical hypothesis and feature correlation are carried out to find out the best suitable model and to capture the relationship between COVID-19 cases with other factors. Also, multivariate analysis using stacked LSTM model is done by training different countries/regions/provinces data and the prediction for Indian COVID-19 cases for 90 days ahead is analyzed.

- The number of confirmed cases at the end of October 2020 is 8.1 million and will cross 9.2 million at the end of November 2020.
- The number of recovered cases is 7.4 million and 9 million at the end of October and November 2020, respectively.

Furthermore, the practical significance for predicting the infected cases is well-established from four perspectives. This includes assessing the pandemic characteristics for the given predicting span and locality, scenario planning in the healthcare sector, industrial sector, government aspects, energy sector and other societal planning in advance, optimization of models and correlating various environmental and societal factors with disease characteristics, and also, supportive aspect in the progress of SDG 3 and 11.

Artificial intelligence, machine learning and deep learning are the key technologies which can help healthcare organizations to support decision making in real-time to control the spread of the pandemic. This study aims to investigate the role of deep learning by analyzing the COVID-19 data to take measures to fight against the pandemic. The forecasted COVID-19 confirmed death and recovered cases that are likely to occur in the near future are predicted using different models. The Stacked LSTM model outperforms the ARIMA, LSTM and Prophet models in predicting the future cases of India and Chennai accurately. Prediction of future COVID-19 cases gives an alert to the people to ensure that they are maintaining social distancing and, by controlling the fast spread of the virus, it is possible to save human lives in the near future.

CRediT authorship contribution statement

Jayanthi Devaraj: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Supervision, Validation, Writing - original draft. **Rajvikram Madurai Elavarasan:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Validation, Writing - original draft. **Rishi Pugazhendhi:** Conceptualization, Formal analysis, Investigation, Visualization, Writing - original draft. **G.M. Shafiullah:** Conceptualization, Validation, Writing - review & editing. **Sumathi Ganesan:** Supervision, Writing - review & editing. **Ajay Kaarthic Jeysree:** Software. **Irfan Ahmad Khan:** Writing - review & editing. **Eklas Hossain:** Writing - review & editing, Funding acquisition.

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.

Acknowledgement

The authors acknowledge the support rendered by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, United States, for providing the COVID-19 real-time dataset for carrying out the research work. The authors would also like to acknowledge the technical expertise provided by the Clean and Resilient Energy Systems (CARES) Laboratory, Texas A&M University,

Galveston, USA.

References

[1] Shereen Muhammad Adnan, Khan Suliman, Kazmi Abeer, Bashir Nadia, Siddique Rabeea. COVID-19 infection: origin, transmission, and characteristics of human coronaviruses. J Adv Res 2020;24:91–8. https://doi.org/10.1016/j.jare.2020.03.005.

[2] Wang Lisheng, Wang Yiru, Ye Dawei, Liu Qingquan. Review of the 2019 novel coronavirus (SARS-CoV-2) based on current evidence. Int J Antimicrob Agents 2020;55(6):105948. https://doi.org/10.1016/j.ijantimicag.2020.105948.

[3] Kumaravel SK, Subramani RK, Jayaraj Sivakumar TK, Madurai Elavarasan R, Manavalanagar Vetrichelvan A, Annam A, Subramaniam U. Investigation on the impacts of COVID-19 quarantine on society and environment: preventive measures and supportive technologies. 3 Biotech 2020;10(9). https://doi.org/10.1007/s13205-020-02382-3.

[4] Zhang Jiancheng, Xie Bing, Hashimoto Kenji. Current status of potential therapeutic candidates for the COVID-19 crisis. Brain Behav Immun 2020;87:59–73. https://doi.org/10.1016/j.bbi.2020.04.046.

[5] Li H, Liu S-M, Yu X-H, Tang S-L, Tang C-K. Coronavirus disease 2019 (COVID-19): current status and future perspectives. Int J Antimicrob Agents 2020;55(5):105951. https://doi.org/10.1016/j.ijantimicag.2020.105951.

[6] Wilder-Smith Annelies, Chiew Calvin J, Lee Vernon J. Can we contain the COVID-19 outbreak with the same measures as for SARS? Lancet Infect Dis 2020;20(5):e102–7. https://doi.org/10.1016/S1473-3099(20)30129-8.

[7] Lai C-C, Shih T-P, Ko W-C, Tang H-J, Hsueh P-R. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): the epidemic and the challenges. Int J Antimicrob Agents 2020;55(3):105924. https://doi.org/10.1016/j.ijantimicag.2020.105924.

[8] El Zowalaty ME, Järhult JD. From SARS to COVID-19: a previously unknown SARS-related coronavirus (SARS-CoV-2) of pandemic potential infecting humans – Call for a One Health approach. One Health 2020;9:100124. https://doi.org/10.1016/j.onehlt.2020.100124.

[9] Mohamed BhedadJamshidi, Jakub Talla, MirhamedMirmozafari, AsalSabet. Artificial intelligence and COVID-19: deep learning approaches for diagnosis and treatment 99; 2020: 1-1. doi:10.1109/ACCESS.2020.3001973.

[10] Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, Eggo RM, Sun F, Jit M, Munday JD, Davies N, Gimma A, van Zandvoort K, Gibbs H, Hellewell J, Jarvis CI, Clifford S, Quilty BJ, Bosse NI, Abbott S, Klepac P, Flasche S. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. Lancet Infect Dis 2020;20(5):553–8. https://doi.org/10.1016/S1473-3099(20)30144-4.

[11] Hellewell J, Abbott S, Gimma A, Bosse NI, Jarvis CI, Russell TW, Munday JD, Kucharski AJ, Edmunds WJ, Funk S, Eggo RM, Sun F, Flasche S, Quilty BJ, Davies N, Liu Y, Clifford S, Klepac P, Jit M, Diamond C, Gibbs H, van Zandvoort K. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. Lancet Global Health 2020;8(4):e488–96. https://doi.org/10.1016/S2214-109X(20)30074-7.

[12] Vaishya Raju, Javaid Mohd, Khan Ibrahim Haleem, Haleem Abid. Artificial Intelligence (AI) applications for COVID-19 pandemic. Diabetes Metab Syndrome 2020;14(4):337–9. https://doi.org/10.1016/j.dsx.2020.04.012.

[13] Naudé W. Artificial Intelligence against COVID-19: An Early Review, IZA Institute of Labor Economics, IZA DP No. 13110, Apr. 2020. [Online]. Available: https://www.iza.org/publications/dp/13110/artificial-intelligence-against-covid-19-an-early-review; 2020.

[14] Carrillo-Larco RM, Castillo-Cara M. Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: an unsupervised machine learning approach. Wellcome Open Res 2020;5:56. https://doi.org/10.12688/wellcomeopenres.15819.1.

[15] Pung R, Chiew CJ, Young BE, Chin S, Chen M-C, Clapham HE, Cook AR, Maurer-Stroh S, Toh MPHS, Poh C, Low M, Lum J, Koh VTJ, Mak TM, Cui L, Lin RVTP, Heng D, Leo Y-S, Lye DC, Lee VJM, Kam K-Q, Kalimuddin S, Tan SY, Loh J, Thoon KC, Vasoo S, Khong WX, Suhaimi N-A, Chan SJH, Zhang E, Oh O, Ty A, Tow C, Chua YX, Chaw WL, Ng Y, Abdul-Rahman F, Sahib S, Zhao Z, Tang C, Low C, Goh EH, Lim G, Hou Y, Roshan I, Tan J, Foo K, Nandar K, Kurupatham L, Chan PP, Raj P, Lin Y, Said Z, Lee A, See C, Markose J, Tan J, Chan G, See W, Peh X, Cai V, Chen WK, Li Z, Soo R, Chow ALP, Wei W, Farwin A, Ang LW. Investigation of three clusters of COVID-19 in Singapore: implications for surveillance and response measures. Lancet 2020;395(10229):1039–46. https://doi.org/10.1016/S0140-6736(20)30528-6.

[16] Madurai Elavarasan R, Pugazhendhi R. Restructured society and environment: a review on potential technological strategies to control the COVID-19 pandemic. Sci Total Environ 2020;725:138858. https://doi.org/10.1016/j.scitotenv.2020.138858.

[17] Fink W, Lipatov V, Konitzer M. Diagnoses by general practitioners: accuracy and reliability. Int J Forecast 2009;25(4):784–93. https://doi.org/10.1016/j.ijforecast.2009.05.023.

[18] Zhu Xianglei, Fu Bofeng, Yang Yaodong, Ma Yu, Hao Jianye, Chen Siqi, Liu Shuang, Li Tiegang, Liu Sen, Guo Weiming, Liao Zhenyu. Attention-based recurrent neural network for influenza epidemic prediction. BMC Bioinf 2019;20(S18). https://doi.org/10.1186/s12859-019-3131-8.

[19] Hewamalage H, Bergmeir C, Bandara K. Recurrent Neural Networks for Time Series Forecasting: Current status and future directions. Int J Forecast 2020;37(1):388–427. https://doi.org/10.1016/j.ijforecast.2020.06.008.

[20] Kırbaş İ, Sözen A, Tuncer AD, Kazancıoğlu FŞ. Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches. Chaos Solitons Fractals 2020;138:110015. https://doi.org/10.1016/j.chaos.2020.110015.

[21] Arora P, Kumar H, Panigrahi BK. Prediction and analysis of COVID-19 positive cases using deep learning models: a descriptive case study of India. Chaos Solitons Fractals 2020;139:110017. https://doi.org/10.1016/j.chaos.2020.110017.

[22] Zeroual A, Harrou F, Dairi A, Sun Y. Deep learning methods for forecasting COVID-19 time-Series data: a Comparative study. Chaos Solitons Fractals 2020;140:110121. https://doi.org/10.1016/j.chaos.2020.110121.

[23] Shahid F, Zameer A, Muneeb M. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. Chaos Solitons Fractals 2020;140:110212. https://doi.org/10.1016/j.chaos.2020.110212.

[24] Vinay Kumar Reddy Chimmula, Lei Zhang. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos Solitons Fractals 135; 2020:109864. doi:10.1016/j.chaos.2020.109864.

[25] Alzahrani Saleh I, Aljamaan Ibrahim A, Al-Fakih Ebrahim A. Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions. J Infect Public Health 2020;13(7):914–9. https://doi.org/10.1016/j.jiph.2020.06.001.

[26] Ogundokun Roseline O, Lukman Adewale F, Kibria Golam BM, Awotunde Joseph B, Aladeitan Benedita B. Predictive modelling of COVID-19 confirmed cases in Nigeria. Infect Disease Model 2020;5:543–8. https://doi.org/10.1016/j.idm.2020.08.003.

[27] Ribeiro MHDM, da Silva RG, Mariani VC, Coelho LDS. Short-term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil. Chaos Solitons Fractals 2020;135:109853. https://doi.org/10.1016/j.chaos.2020.109853.

[28] Tomar A, Gupta N. Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. Sci Total Environ 2020;728:138762. https://doi.org/10.1016/j.scitotenv.2020.138762.

[29] Car Zlatan, Baressi Šegota Sandi, Anđelić Nikola, Lorencin Ivan, Mrzljak Vedran. Modeling the spread of COVID-19 infection using a multilayer perceptron. Comput Math Methods Med 2020;2020:1–10. https://doi.org/10.1155/2020/5714714.

[30] Shastri Sourabh, Singh Kuljeet, Kumar Sachin, Kour Paramjit, Mansotra Vibhakar. Time series forecasting of Covid-19 using deep learning models: India-USA comparative case study. Chaos Solitons Fractals 2020;140:110227. https://doi.org/10.1016/j.chaos.2020.110227.

[31] Hawas M. Generated time-series prediction data of COVID-19′s daily infections in Brazil by using recurrent neural networks. Data Brief 2020;32:106175. https://doi.org/10.1016/j.dib.2020.106175.

[32] Vasilis Papastefanopoulos, Pantelis Linardatos, Sotiris Kotsiantis. COVID-19: a comparison of time series methods to forecast percentage of active cases per population. Appl Sci 10; 2020: 3880. doi:10.3390/app10113880.

[33] Jiménez F, Palma J, Sánchez G, Marín D, Francisco Palacios MD, Lucía López MD. Feature selection based multivariate time series forecasting: an application to antibiotic resistance outbreaks prediction. Artif Intell Med 2020;104:101818. https://doi.org/10.1016/j.artmed.2020.101818.

[34] Yang Jie, Ma Jun. Feed-forward neural network training using sparse representation. Expert Syst Appl 2019;116:255–64. https://doi.org/10.1016/j.eswa.2018.08.038.

[35] Wang Kang, Li Kenli, Zhou Liqian, Hu Yikun, Cheng Zhongyao, Liu Jing, Chen Cen. Multiple convolutional neural networks for multivariate time series prediction. Neurocomputing 2019;360:107–19. https://doi.org/10.1016/j.neucom.2019.05.023.

[36] Lafta R, Zhang J, Tao X, Li Y, Abbas W, Luo Y, et al. A fast Fourier transform-coupled machine learning-based ensemble model for disease risk prediction using a real-life dataset; 2017: 654–670. doi:10.1007/978-3-319-57454-7_51.

[37] Lei L, Zhou Y, Zhai J, Zhang L, Fang Z, He P, et al. An effective patient representation learning for time-series prediction tasks based on EHRs. IEEE Int Conf Bioinf Biomed 2018;2018:885–92. https://doi.org/10.1109/BIBM.2018.8621542.

[38] Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Sci Rep 2018;8(1):6085. https://doi.org/10.1038/s41598-018-24271-9.

[39] Parmezan ARS, Souza VMA, Batista GEAPA. Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model. Inf Sci 2019;484:302–37. https://doi.org/10.1016/j.ins.2019.01.076.

[40] Rizk Yara, Awad Mariette. On extreme learning machines in sequential and time series prediction: a non-iterative and approximate training algorithm for recurrent neural networks. Neurocomputing 2019;325:1–19. https://doi.org/10.1016/j.neucom.2018.09.012.

[41] Yadav Anita, Jha C K, Sharan Aditi. Optimizing LSTM for time series prediction in Indian stock market. Proc Comput Sci 2020;167:2091–100. https://doi.org/10.1016/j.procs.2020.03.257.

[42] Fang X, Yuan Z. Performance enhancing techniques for deep learning models in time series forecasting. Eng Appl Artif Intell 2019;85:533–42. https://doi.org/10.1016/j.engappai.2019.07.011.

[43] Taylor SJ, Letham B. Forecasting at scale. PeerJ Preprints 2017;2017. https://doi.org/10.7287/peerj.preprints.3190v2.

[44] Zhao Naizhuo, Liu Ying, Vanos Jennifer K, Cao Guofeng. Day-of-week and seasonal patterns of PM2.5 concentrations over the United States: time-series analyses using the Prophet procedure. Atmos Environ 2018;192:116–27. https://doi.org/10.1016/j.atmosenv.2018.08.050.

[45] Papacharalampous GA, Tyralis H. Evaluation of random forests and Prophet for daily streamflow forecasting. Adv Geosci 2018;45:201–8. https://doi.org/10.5194/adgeo-45-201-2018.
[46] CSSE, COVID 19 time series dataset of confirmed, death and recovered cases by the Center for Systems and Engineering (CSSE) at Johns Hopkins University, United States. Available: https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases; 2020 [accessed August 22, 2020].
[47] Ramírez-Gallego S, Krawczyk B, García S, Woźniak M, Herrera F. A survey on data preprocessing for data stream mining: current status and future directions. Neurocomputing 2017;239:39–57. https://doi.org/10.1016/j.neucom.2017.01.078.
[48] Yang Lingzhi, Ban Xiaojuan, Chen Zhe, Guo Hongyue. A new data preprocessing technique based on feature extraction and clustering for complex discrete temperature data. Proc Comput Sci 2018;129:78–80. https://doi.org/10.1016/j.procs.2018.03.050.
[49] Czarnowski I, Jedrzejowicz P. Data reduction algorithm for machine learning and data mining. In New Frontiers in Applied Artificial Intelligence (pp. 276–285). Springer Berlin Heidelberg; 2008. doi:10.1007/978-3-540-69052-8_29.
[50] Vantuch Tomas, Snasel Vaclav, Zelinka Ivan. Dimensionality reduction method's comparison based on statistical dependencies. Proc Comput Sci 2016;83:1025–31. https://doi.org/10.1016/j.procs.2016.04.218.
[51] Kaushik S, Choudhury A, Sheron PK, Dasgupta N, Natarajan S, Pickett LA, et al. AI in healthcare: time-series forecasting using statistical, neural, and ensemble architectures. Front Big Data 3; 2020. doi:10.3389/fdata.2020.00004.
[52] Martin Hagan, Neural Network Design, http://hagan.okstate.edu/NNDesign.pdf.
[53] COVID 19 India. Chennai dataset, https://www.covid19india.org/state/TN; 2020 [accessed 23 August 2020].
[54] Shafiullah GM. Hybrid renewable energy integration (HREI) system for subtropical climate in Central Queensland, Australia. Renew Energy 2016;96:1034–53. https://doi.org/10.1016/j.renene.2016.04.101.
[55] World Weather Page: https://www.weather2visit.com/ (accessed on 10/11/2020).
[56] Jiang X, Coffee M, Bari A, Wang J, Jiang X, Huang J, et al. Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. Comput Mater Continua 62(3); 2020: 537–551. doi:10.32604/cmc.2020.010691.
[57] Wang Lishi, Li Jing, Guo Sumin, Xie Ning, Yao Lan, Cao Yanhong, Day Sara W, Howard Scott C, Graff J Carolyn, Gu Tianshu, Ji Jiafu, Gu Weikuan, Sun Dianjun. Real-time estimation and prediction of mortality caused by COVID-19 with patient information based algorithm. Sci Total Environ 2020;727:138394. https://doi.org/10.1016/j.scitotenv.2020.138394.
[58] Madurai Elavarasan R, Shafiullah GM, Raju K, Mudgal V, Arif MT, Jamal T, Subramanian S, Sriraja Balaguru VS, Reddy KS, Subramaniam U. COVID-19: Impact analysis and recommendations for power sector operation. Appl Energy 2020;279:115739. https://doi.org/10.1016/j.apenergy.2020.115739.
[59] Diginomica. How Canadian AI start-up BlueDot spotted Coronavirus before anyone else had a clue. https://diginomica.com/how-canadian-ai-start-bluedot-spotted-coronavirus-anyone-else-had-clue; 2020 [accessed 24 August 2020].
[60] Sobral Marcos Felipe Falcão, Duarte Gisleia Benini, da Penha Sobral Ana Iza Gomes, Marinho Marcelo Luiz Monteiro, de Souza Melo André. Association between climate variables and global transmission oF SARS-CoV-2. Sci Total Environ 2020;729:138997. https://doi.org/10.1016/j.scitotenv.2020.138997.
[61] Sun Zhibin, Zhang Hui, Yang Yifei, Wan Hua, Wang Yixiang. Impacts of geographic factors and population density on the COVID-19 spreading under the lockdown policies of China. Sci Total Environ 2020;746:141347. https://doi.org/10.1016/j.scitotenv.2020.141347.