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

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Comparative analysis of Gated Recurrent Units (GRU), long Short-Term memory (LSTM) cells, autoregressive Integrated moving average (ARIMA), seasonal autoregressive Integrated moving average (SARIMA) for forecasting COVID-19 trends

K.E. ArunKumar a,*, Dinesh V. Kalaga b,*, Ch. Mohan Sai Kumar c, Masahiro Kawaji b, Timothy M. Brenza a,*

a Department of Chemical and Biological Engineering, South Dakota Mines, An Engineering, Science and Technology University, Rapid City, SD 57701, USA
b Mechanical Engineering Department, City College of New York, New York, NY 10031, USA
c Process Chemistry and Technology, CSIR-Central Institute of Medicinal and Aromatic Plants, Lucknow, UP 226015, India

Received 19 November 2020; revised 25 November 2021; accepted 2 January 2022
Available online 06 January 2022

Abstract Several machine learning and deep learning models were reported in the literature to forecast COVID-19 but there is no comprehensive report on the comparison between statistical models and deep learning models. The present work reports a comparative time-series analysis of deep learning techniques (Recurrent Neural Networks with GRU and LSTM cells) and statistical techniques (ARIMA and SARIMA) to forecast the country-wise cumulative confirmed, recovered, and deaths. The Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM) cells based on Recurrent Neural Networks (RNN), ARIMA and SARIMA models were trained, tested, and optimized to forecast the trends of the COVID-19. We deployed python to optimize the parameters of ARIMA which include (p, d, q) representing autoregressive and moving average terms and parameters of SARIMA model include additional seasonal terms which are denoted by (P, D, Q). Similarly, for LSTM and GRU based RNN models’ parameters (number of layers, hidden size, learning rate and number of epochs) are optimized by deploying PyTorch machine learning framework. The best model was chosen based on the lowest Mean Square Error (MSE) and Root Mean Squared Error (RMSE) values. For most of the time-series data of the countries, deep learning-based models LSTM and GRU outperformed statistical ARIMA and SARIMA models, with an RMSE values...
1. Introduction

A strain of coronaviruses known as Severe Acute Respiratory Syndrome CoronaVirus-2 (SARS-CoV-2) causes coronavirus disease (COVID-19) which is a respiratory illness and was first identified on 31st December 2019 in Wuhan China. Soon after, World Health Organization (WHO) declared COVID-19 as pandemic on 11th March 2020 [1]. Ever since COVID-19 was declared as a pandemic, many countries implemented preventive measures such as social distancing, rapid COVID-19 testing, vaccination drives, restricted mobility of the civilians, and lockdowns to curb the spread of COVID-19 pandemic. However, COVID-19 is spreading at a faster pace eventually affecting the global economy and human health. The factor causing the spread of COVID-19 includes population density, total population, and lifestyle, etc. As of today (1st August 2021), there have been about cumulative confirmed cases of 206 M and about total fatalities of 4.34 M across the globe. Most of these cumulative cases are concentrated in a very few countries such as the USA, India, Brazil, Russia, South Africa, Mexico, Peru, Chile, United Kingdom, Iran [2]. As COVID-19 has adversely impacting the public health and global economy, it is important to get an estimate of what would be the upcoming number of daily new cases, total fatalities, total confirmed cases, and total recovered cases. The prior knowledge on the upcoming cases can help healthcare systems and governments to prepare for the forecasted number of cases. Due to the availability of abundant data, statistical machine learning and deep learning techniques have gained immense interest of researchers for forecasting the trends based on the accumulated data over a period of time. These techniques are advantageous for obtaining relationships with future time from the data without defining them prior [3].

Machine learning techniques such as linear regression, Support Vector Regression (SVR), to name few, are being implemented for forecasting the COVID-19 pandemic by several researchers. For example, linear regression and multiple linear regression analysis were conducted by Ghosal et al. [4] to predict the number of deaths in India for upcoming 6 weeks. They reported that COVID-19 caused deaths will be doubled in India if COVID-19 preventive measures are unchanged or not implemented. Prabat and Chakraborty [5] have forecasted the COVID-19 cases in India for the next 60 days based on the reported time-series data. They developed support vector regression models for forecasting daily new cases with an accuracy of 87% and for predicting the total number of fatalities, total number of recovered cases, and total number of confirmed cases with an accuracy of 97%. Maleki et al.[6] have developed Two-Piece Mixture Scale Normal Distribution Auto Regressive (TP-SMN-AR) models to forecast the cumulative confirmed and cumulative recovered COVID-19 cases. Further, the proposed TP-SMN-AR models performed well. Benvenuto et al. [7] have developed ARIMA models to predict the future trends of prevalence and incidence of COVID-19 pandemic. They have also briefly reviewed the applications of ARIMA models. Ribeiro et al. [8] have reported a comparative time-series analysis of Brazil’s data. They reported a short-term forecast using models such as Stacking-ensemble learning, Ridge Regression, Cubist Regression, Random Forest, ARIMA, and SVR. Their study reports that the SVR model’s performance is better than ARIMA model. Kumar et al. [9] have forecasted the trajectories of the COVID-19 pandemic for the top 15 countries in the month of April 2020. The developed ARIMA model was successfully able to predict the COVID-19 pandemic in countries such as Iran, Spain, Italy, Spain, and France. Ardabili et al. [10] have employed a multi-layer perceptron model and adaptive network-based fuzzy interface system for predicting the COVID-19 trends and the developed models were found to give a promising predictions of the COVID-19 data. Their investigation recommended the development of country-specific machine learning models due to the presence of fundamental differences between time-series-data of countries. The following paragraph briefly summarizes the reported work on the deep learning-based forecast of COVID-19 cases.

Chimmula and Zhang [11] have developed a state-of-art deep learning model using LSTM network to estimate the end time of the COVID-19 pandemic and their LSTM network predicted the COVID-19 pandemic in Canada will end in about a period of 3 months. Salgotra et al. [12] developed genetic algorithm based models for forecasting the cumulative (confirmed, recovered, deaths) cases in highly affected states of India as well as for total India. Their models were less sensitive to the variables and have high reliability in terms of predicting the cumulative confirmed and cumulative death cases. Qi et al. [13] have developed an adaptive model for drawing a correlation between daily average temperature, relative humidity with daily COVID-19 cases in the various provinces of China. Their investigation revealed a negative correlation between the daily temperature, relative humidity, and daily COVID-19 cases, meaning, as the daily temperature and relative humidity increases the number of COVID-19 cases decreases. However, authors found that there is no clear trend for throughout mainland China. In a study, Ismail et al. [14] have forecasted COVID-19 cases in countries such as Denmark, Belgium, Germany, France, UK, Switzerland, and Turkey using LSTM, ARIMA, Non-Linear Auto regressive Neural Networks (NARNN). Their study revealed that the LSTM model offered the lowest RMSE compared to the other models. In a study, Dutta et al. [15] have predicted the confirmed, released, negative, and death cases of COVID-19 pandemic using LSTM, GRU, and RNN models. Their study revealed that a hybrid model (LSTM-RNN) performed better than the individual models. Tomar et al. [16] have developed a curve fitting based
models and LSTM models for forecasting the COVID-19 cases in India for the duration of the upcoming 30 days. They also investigated the effect of preventive measures such as lockdowns, social distancing, etc. on the spread of the COVID-19. Their study revealed that the number of daily cases has decreased according to LSTM and curve-fitting methods. Mehdi et al. [17] conducted a comparative analysis of RNN, LSTM, SARIMA and holt winter’s exponential smoothing models on the COVID-19 data of Iran. They found that the LSTM models perform better than other models for forecasting Iran’s data with low evaluation metric values. Gaetano et al. [18] have conducted a comparative analysis of COVID-19 hospitalization data using ARIMA, exponential smoothing (ETS), the neural network autoregression (NNAR) models. The trigonometric exponential smoothing state space model with Box–Cox transformation, ARMA trend and seasonal components. They reported that the feasible hybrid combination of these models performed better in terms of forecasting the number of patients admitted with mild symptoms and number of patients admitted to intensive care units (ICU) [18]. Similarly, Maher et al. [19] have developed a hybrid dynamic model based on Susceptible Exposed Infective Recovered Death (SEIRD) with ARIMA corrections that can perform short- and long-term forecasting. Soudh et al. [20] have also performed a comparative analysis of COVID-19 data by deploying ARIMA, Adaptive neuro-fuzzy inference system (ANFIS), and Multilayer perceptron (MLP), artificial neural network (ANN) and Bidirectional long short-term memory (LSTM) models. They have reported that these models are capable of capturing the dissimilarities in COVID-19 spread across various regions or populations. Oluwatamilore et al. [20] have developed a sluggish state based neural networks to forecast COVID-19. This developed multi-recurrent network (MRN) was used as an alternative LSTM models. In another study, Joshua et al. [21] have developed a simple ARIMA model for forecasting the COVID-19 vaccination process. They have also developed an ARIMA based probability distribution curve rates to classify active cases.

Similarly, multiple historical ARIMA models were reported by Jian Sun [22] for 13 weeks ahead forecast of COVID-19 pandemic in Alberta, Canada. He reported that the average incidents forecasted with historical ARIMA models were lower than those with general ARIMA models.

Roy et al. [23] have used the ARIMA models for forecasting COVID-19 daily cases of Indian states that have high incidences. They have also performed out of sample forecasting using MAE and RMSE as evaluation metrics. The following Table 1 summarizes the recent work on the forecasting COVID-19 using various machine learning and deep learning models.

Based on the reports in the literature, there are several studies on forecasting the COVID-19 pandemic, but they are restricted to either a few countries, few states of a country, or projected a short-term forecast. These models were developed on very short length of the reported data. There are not many reports on the comparative analysis on the deep learning models and statistical methods for COVID-19 data. Moreover, the reported comparative analysis in the literature did not include the comparison between ARIMA, SARIMA, GRU-RNN, LSTM-RNN models for top-10 highly affected countries. Therefore, there is still room to develop country-specific GRU, LSTM, ARIMA and SARIMA models to predict the 60-day ahead forecast of COVID-19 trends in the top 10 countries that are highly impacted by the pandemic and provide a comparative analysis on these models. Hence, the main objective of the work is to predict the future trends of the cumulative confirmed cases, cumulative recovered cases, and cumulative fatalities of the top-10 countries using RNN based GRUs, LSTM cells, ARIMA and SARIMA and to illustrate the relative performance of the ARIMA, SARIMA, LSTM-RNN and GRU-RNN models. The models were trained and tested on training and testing data respectively, these validated models were then used to forecast the trends of COVID-19 cases for 60 days in the selected 10 countries. The time-series data, reported up until 22nd of June 2021 was used for developing the models was obtained from John Hopkins website [24]. The statistical models ARIMA and SARIMA were developed by deploying python statistical libraries on a local personal computer and for the deep learning models, the python-based deep learning library PyTorch was deployed on Graphical Processing Units (GPU’s) of google COLAB cloud computing platform.

2. Mathematical modelling

Deep learning methods are widely used to forecast the non-linear datasets such as weather data, stock prices [35], electrocardiogram (ECG) recordings [36] and crude oil prices [37] etc. There are two types of widely used deep learning techniques such as Feed Forward Neural Networks (FFNNs) and Recurrent Neural Networks (RNNs) but FFNNs are not suitable for forecasting the time-series data as they are not capable of capturing trends in the time-series data. On the other hand, RNN’s are robust artificial neural networks that use existing time-series data to predict the future data over a specific length of time. RNN’s internal memory makes it a very promising technique because they support in remembering the important features of the input sequential data which allows them to accurately predict the future. The major difference between the FFNNs and RNNs is that in FFNNs the information flows strictly in one direction from layer to layer, whereas in RNNs, the input from the present time-step and the output from the previous time-stamp will be fed into RNN cells so that the current state of the model is influenced by its previous states.

The single RNN cell is mathematically represented by following equation:

$$h_t = \tanh(W[h_{t-1}, x_t] + b)$$ (1)

Where, b is the bias matrix, W is the weight matrix, $$h_{t-1}$$ and $$h_t$$ are hidden state at current time-step and previous time-step, respectively. The computation in the RNNs cells will be the same using the weights, biases, and activation functions for each element of the input sequence as shown in Fig. 1A. Fundamentally a neuron of the RNN model uses a single hyperbolic tangent function in which the $$h_t$$ and $$x_t$$ are combined followed by multiplication with some weight matrix, and then a bias is added. Then this value is passed through hyperbolic tangent function which gives back $$h_t$$. Hyperbolic tan function ($$\tanh$$) transforms the input data so that the values lie in the range of $$-1$$ to $$+1$$. The following equation (Eq. (2)) is a mathematical representation of the sigmoid function
which is used to introduce non-linearity to the artificial neural networks, which is used at each time-step to update the output of the RNN cell.

\[ f(x) = \frac{1}{1 + e^{-x}} \]  

RNN models can recollect the recent information and cannot recollect the earlier information that is far away in time from the current time-step. RNN cannot be trained for long input sequences due to the vanishing gradient problem. Moreover, the RNN architecture has shorter memory.

| Authors                  | Models and parameters reported in Literature | Remarks                                           |
|--------------------------|---------------------------------------------|--------------------------------------------------|
| Hitesh Tandon et al. [25] | ARIMA: p=2, d=2, Q=NA, SARIMA: (p, d, Q)=(2,2,1) | Single exponential moving average, S-Curve trend model. Study conducted on one country and did not include deep learning models. |
| Koyuncu et al. [26]      | ARIMA: p=NA, d=NA, Q=NA, SARIMA: (2,2,1) | Exponential smoothing. Forecasted the impact of COVID-19 on ISL container throughput index. This study reported the comparison between LSTM and LSTM-Markov model but did not discuss the parameters of these models. The proposed hybrid CNN-LSTM model performed better than ARIMA or LSTM models. This study also lacks the information on the parameters of the models. This study did not provide suggestion on which model performed better and recommended Bayesian inference framework for COVID-19 forecast studies. |
| Ruifang Ma et al. [27]   | ARIMA: p=NA, d=NA, Q=NA, SARIMA: (2,2,1) | LSTM, LSTM-Markov hybrid model |
| Shwet Ketu et al. [28]   | ARIMA: p=NA, d=NA, Q=NA, SARIMA: (2,2,1) | CNN-LSTM, LSTM, ARIMA |
| Chandra et al. [29]      | ARIMA: p=NA, d=NA, Q=NA, SARIMA: (2,2,1) | LSTM, Bidirectional-LSTM, Encoder Decoder-LSTM |
| Demir [30]               | ARIMA: p=NA, d=NA, Q=NA, SARIMA: (2,2,1) | Neural Network Nonlinear Autoregressive (NNAR) model, SARIMA-NNAR. |
| Saina et al. [31]        | ARIMA: p=3, d=2, Q=3, SARIMA: (3,2,2) | Holt-Winter Exponential smoothing additive (HWESA) model. |
| Dairi et al. [32]        | ARIMA: p=NA, d=NA, Q=NA, SARIMA: (5,3) | LSTM-CNN, Generative Adversarial Network (GAN-GRU) hybrid model, Restricted Boltzmann Machine (RBM) |
| Nahla et al. [33]        | ARIMA: p=NA, d=NA, Q=NA, SARIMA: (2,2) | CNN, Multivariate CNN |
| Nabi et al. [34]         | ARIMA: p=NA, d=NA, Q=NA, SARIMA: Brazil (10), Russia (3) | CNN, Multivariate CNN |

Table 1 Models reported in literature for forecasting COVID-19 pandemic.
to remember the features and it faces the problem of vanishing and exploding gradients. Therefore, RNNs are usually combined with LSTM cells and GRUs to overcome these drawbacks.

2.1. Long short-term memory (LSTM)

To overcome the vanishing gradients and exploding gradients problem, Hochreiter and Schmidhuber [38] have proposed
LSTM. The advantage of LSTM is that it has a cell state which stores and converts the input cell memory to the output cell state. Fig. 1B shows the general architecture of the LSTM cell, which mainly consists of input gate, output gate, forget gate, and update gate. The forget gate decides what to forget from the information received through previous memory units, input gate decides what information to accept into the neuron, output gate generates the new long-term memory, and the update gate updates the cell. These four components work and interact in a specific manner, as it accepts the short-term, long-term memories, input sequence at a given timestamp and creates a new long-term, short-term memories, and output sequences at a given time-step. The following equation is a mathematical expression of the input gate which determines what information must be transferred to the cell:

\[ i_t = \sigma(W_{x} * [h_{t-1}, x_t] + b_i) \]  

Whereas the following equation is a mathematical expression of forget gate which determines what information to be neglected:

\[ f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f) \]  

The update gate updates the cell state which is mathematically expressed by the following equations:

\[ c_t = \tanh(W_c * [h_{t-1}, x_t] + b_c) \]  
\[ c_t = f_t * c_{t-1} + i_t * c_t \]  

The output gate is responsible for the updating the output and is given by following equation, the output gate is also responsible for updating the hidden layer of the previous time-step.

\[ o_t = \sigma(W_o * [h_{t-1}, x_t] + b_o) \]  
\[ h_t = o_t * \tanh(c_t) \]  

2.2. Gated Recurrent unit (GRU)

Chung et al. [39] proposed a simplified version of the LSTM cell which is called as Gated Recurrent Units (GRUs), it requires the less training time with improved network performance (Fig. 1C). In terms of operation, GRU and LSTM works similarly but GRU cell uses one hidden state that merges the forget gate and the input gate into a single update gate. Moreover, GRU combines the hidden and cell states into one state. Therefore, the total number of gates in GRU is half of the total number of gates in LSTM making GRU popular and a shortened variant of LSTM cell. The two gates of GRU are update gate and reset gate. The following equation represents the hidden state of the GRU:

\[ h_t = (1 - z_t) * h_{t-1} + z_t * h_t \]  

The following equation represents the update gate and determines how much of the GRU unit get updated:

\[ z_t = \sigma(W_z * [h_{t-1}, x_t]) \]  

The reset gate is given by the following equation:

\[ r_t = \sigma(W_r * [h_{t-1}, x_t]) \]  

The hyperbolic tangent function of the reset gate is called as new remember a gate which is described by the following function.

\[ h_t = \tanh(W * [r_t * h_{t-1}, x_t]) \]  

2.3. Auto regressive integrated moving average (ARIMA)

ARIMA (p, d, q) can be used for forecasting of stationary timeseries data, the model parameters (p, d, q) can be defined as follows, the Auto Regression (AR) term (p), Moving Average (MA) term (q), ARIMA models are build up on integrating these two parts of the models using differencing term (d or I). The AR term is the regression of a specific variable against itself to forecast the variable of interest. Whereas the MA term is based on the error terms of the forecast at a previous time-step to forecast a variable at later time-step. The following equation (Eq. (13)) generalizes the pth order AR model and qth order MA model respectively.

\[ y_t = C + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \epsilon_t \]  

Where C is intercept, \( \phi(i = 1, 2, \ldots p) \) is an auto regressive parameter, \( y_t \) is current time-series value \( y_{t-1}, y_{t-2}, y_{t-p} \) are past values and \( \epsilon_t = y_t - y_{t-1} \)

2.4. Seasonal auto regressive integrated moving average (SARIMA)

SARIMA model includes both ARIMA parameters (p, d, q) and seasonal terms (P, D, Q), where P is seasonal AR term, D is seasonal differencing term and Q is seasonal moving average term. The SARIMA model is mathematical represented as follows:

\[ \hat{y}_t = \Theta_q(B^P) \Theta_P(B) (1-B)^D y_t \]  

Where \( \gamma \) is the non-stationary time-series, \( \psi \) is the Gaussian white noise process, \( \hat{y}_t \) is seasonal moving average polynomial, \( \Theta_q(B^P) \) is a seasonal moving average polynomial. Where B is a backshift operator. Further mathematical details of ARIMA and SARIMA models was described elsewhere [40].

2.5. Simulation methodology

The models ARIMA and SARIMA were developed using the pyramid arima library of python and the deep learning models were written using the PyTorch package of python on google colab. The ARIMA and SARIMA models were performed in a windows 10 PC with 8 GB RAM and the processor was Intel i7. The average computation time taken for training and forecasting of ARIMA and SARIMA models on the local PC is less than 6 s. The configuration of Google Collaboratory is GPC - Nvidiak80/T4, memory of the GPU of 16 GB, the GPUTC memory clock id 1.59 GHz, available ram is 26.75 GB and number of CPU cores were 2. The average time taken for training LSTM and GRU models on Google Collaboratory’s GPU is 60 ± 10 min and for forecasting the data
using LSTM and GRU models on GPU is 45 ± 10 min. For each time-series data (cumulative confirmed, recovered, and deaths) of each country, the process was repeated, and the model's prediction was validated against the corresponding test data. Optimization of the parameters of both statistical models (ARIMA and SARIMA) and deep learning models (GRU-RNN and LSTM-RNN) is crucial for achieving the best predictions and forecasts. To avoid the influence of the outliers and other fluctuation of the time-series data on the models and to capture the trends in the reported time-series data, we normalized the data before developing the ARIMA, SARIMA and recurrent neural networks based on GRU and LSTM. Minmax scaler was used to normalize the data and the following equation (Eq. (15)) is a mathematical representation of minmax scaler:

\[ x_n = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \]  

(15)

Where, \( x \) is original time-series data, \( x_n \) is normalized time-series data, \( x_{\text{min}} \) is minimum value in the timeseries, and \( x_{\text{max}} \) is the maximum value of the timeseries. The advantage of using minmax scaler is that it transforms data such that the minimum and maximum value of the feature lies between 0 and 1. Moreover, it retains the shape of the original distribution of the data, and it does not alter the information that is embedded within the original data. Followed by data normalization, the data set was divided into testing and training data sets, the test data is comprised of the last 14 data points and the training data set was divided into testing and training data sets, the test data is comprised of the last 14 data points and the training set contains the entire dataset excluding the testing data.

The evaluation metrics used to evaluate the performance of the proposed models were Mean Square Error (MSE) and Root Mean Square Error (RMSE) which were mathematically defined by following equations Eq. (16) and Eq. (17):

\[ \text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]  

(16)

\[ \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \]  

(17)

Where, \( \hat{y} \) is model predicted value, \( y_i \) is actual value.

The hyperparameters of deep learning models include nodes (10, 100, 200, 300), number of hidden layers (1, 2, 3, 4, 5) and learning rate (0.1, 0.01, 0.001, 0.0001). The values in the parenthesis are the range of each hyperparameter that was studied during the optimization of each model for country-specific data. Adam optimizer was used to iteratively optimize the weights of the network using mean squared error (MSE) as loss function. The Adam optimizer was selected because the COVID-19 data is of some countries is sparse and Adam offers adaptive learning rate which is perfect for such data; therefore, we do not have to optimize learning rate to a great extent when Adam is used. Further, the training data was divided into several data sequences each of length 30 (i.e., day 1–day 30) and each of the sequences is fed to the RNN model to predict the 31st data point. In the following step, the next sequence of the data (day 2–day 31) is fed to RNN to predict the 32nd day data point. Similarly, all the sequences were fed to RNN to complete one set of 60 predictions, and which also completes one epoch. The models were training for approximately 5,000 epochs and the best epoch was identified and used for the final forecast. The best epoch corresponds to the global minima of the loss function is identified as the best epoch. Initial optimization of the hyper parameters was done based on the USA data, manual hyperparameter optimization was performed by running several simulations with all possible combinations of the hyperparameters. The optimized number of the hidden layers, neurons for each layer and learning rate needed for LSTM and GRU was obtained are worked reasonably well. For the optimization of hyperparameters for rest of the countries, we used convergence plots while running the iterations and manually stopped when we reached the lowest MSE and RMSE values.

The parameters of ARIMA are \((p,d,q)\) and SARIMA \((p,d,q)(P,D,Q)_m\) were optimized using auto arima module of pyramid arima library of python [41]. Second-order differencing was required for most of the countries’ cumulative cases. ACF plot of stationary time-series was used to get a basic idea on whether AR terms or MA terms will fit to the data to deliver a superior model. Selecting the best parameter \((p, d, q)\) To select the proper combination of the model parameter values we performed a grid search using pmdarima (Pyramid ARIMA) library available in statsmodels. (a python module). The pmdarima uses AIC as an evaluation metric to choose the best model from various ARIMA and SARIMA models. The SARIMA models were developed by using stepwise parameter selection identify the best combination by setting the seasonality to “True” during the grid search. Since the cumulative COVID-19 cases are of only few months, the parameter that represents seasonality \((m)\) was assigned to 3, 7, 12. Our data analysis showed that seasonality terms varied from country to country. The model with the best seasonal term was identified using information criteria (AIC and BIC). Zohair Malki et al. [42]. Further details on development of ARIMA and SARIMA models is described in detail by K E ArunKumar et al. [40].

3. Results and discussion

The forecasted trends of cumulative confirmed, cumulative recovered, and cumulative fatalities of the COVID-19 cases in 10 countries using GRU-RNN, LSTM-RNN, ARIMA, and SARIMA models are described in this section. The countries based on cumulative confirmed cases data, as of June 22nd, 2021, includes the USA, Brazil, India, Russia, South Africa, Mexico, Peru, Chile, United Kingdom (UK) and Iran. Each country contains three time-series data sets: cumulative confirmed cases, cumulative recovered cases, and cumulative fatalities. Each of these time-series data sets were used for forecasting the trends by optimized GRU-RNN, LSTM-RNN, ARIMA, and SARIMA models. The forecast was done for each of the three time-series data of individual countries by independently feeding the time-series data into the optimized GRU-RNN, LSTM-RNN, ARIMA, and SARIMA models, therefore a total of 120 simulations were performed.

3.1. Cumulative confirmed cases

For country-specific time series data, we proposed customized models and compared the forecast of ARIMA, SARIMA, RNN-GRU, RNN-LSTM models. 60-day forecast of the cumulative confirmed cases obtained from the ARIMA, SARIMA, RNN-LSTM, and RNN-GRU models for all 10 coun-
| Country   | Models used                                                                 | Epochs       | Hidden size       | No.of layers | Learning rate | MSE          | RMSE          |
|-----------|------------------------------------------------------------------------------|--------------|-------------------|--------------|---------------|--------------|---------------|
| USA       | GRULSTMARIMA(5,2,2) SARIMA(4,2,0) (2,0,0,7)                                  | 5.18E + 026.50E + 02 | 3.00E + 023.00E + 00 | 2.00E + 002.00E + 00 | 1.00E-051.00E-05 | 1.20E + 081.48E + 05 | 1.10E + 041.21E + 044.07E + 04 |
| Brazil    | GRULSTMARIMA(1,2,4) SARIMA(0,2,1)(1,0,1,7)                                  | 1.00E + 031.00E + 03 | 3.00E + 023.00E + 00 | 2.00E + 002.00E + 00 | 1.00E-051.00E-05 | 1.44E + 099.38E + 09 | 1.20E + 059.68E + 041.11E + 04 |
| India     | GRULSTMARIMA(1,2,1) SARIMA (1, 2, 0)(1, 0, 0, 7)                            | 8.94E + 021.04E + 03 | 3.00E + 023.00E + 00 | 2.00E + 002.00E + 00 | 1.00E-051.00E-05 | 5.93E + 097.61E + 09 | 7.70E + 048.72E + 044.74E + 04 |
| Russia    | GRULSTMARIMA(3,2,6) SARIMA (1, 2, 0)(1, 0, 1, 7)                            | 1.75E + 031.19E + 03 | 3.00E + 023.00E + 00 | 2.00E + 002.00E + 00 | 1.00E-051.00E-05 | 6.66E + 087.46E + 09 | 2.58E + 048.64E + 033.89E + 04 |
| South Africa | GRULSTMARIMA(4,2,6) SARIMA(4,2,0)(2,0,0,7)                                  | 8.45E + 028.41E + 02 | 3.00E + 023.00E + 00 | 2.00E + 002.00E + 00 | 1.00E-051.00E-05 | 1.23E + 081.43E + 05 | 1.11E + 041.20E + 044.42E + 04 |
| Mexico    | GRULSTMARIMA(4,2,5) SARIMA(0,2,1)(1,0,1,7)                                  | 8.50E + 028.85E + 02 | 3.00E + 023.00E + 00 | 2.00E + 002.00E + 00 | 1.00E-051.00E-05 | 3.35E + 072.78E + 06 | 5.79E + 035.27E + 036.37E + 03 |
| Peru      | GRULSTMARIMA(1,2,1) SARIMA (5, 2, 1)(2, 0, 0, 7)                            | 1.50E + 031.40E + 03 | 3.00E + 023.00E + 00 | 2.00E + 002.00E + 00 | 1.00E-051.00E-05 | 7.48E + 075.43E + 07 | 8.65E + 037.37E + 036.64E + 03 |
| Chile     | GRULSTMARIMA(5,2,5) SARIMA (3, 2, 0)(2, 0, 1, 7)                            | 2.00E + 039.68E + 02 | 3.00E + 023.00E + 00 | 2.00E + 002.00E + 00 | 1.00E-051.00E-05 | 1.06E + 085.64E + 05 | 1.03E + 047.51E + 031.89E + 04 |
| UK        | GRULSTMARIMA(6,2,4) SARIMA (6, 2, 0)(0, 0, 2, 7)                            | 1.20E + 031.75E + 03 | 3.00E + 023.00E + 00 | 2.00E + 002.00E + 00 | 1.00E-051.00E-05 | 6.62E + 087.06E + 05 | 2.57E + 042.66E + 041.71E + 04 |
| Iran      | GRULSTMARIMA(2,2,3) SARIMA(1,2,0)(1,0,1,7)                                  | 5.18E + 026.50E + 02 | 3.00E + 023.00E + 00 | 2.00E + 002.00E + 00 | 1.00E-051.00E-05 | 1.20E + 081.20E + 083.94E + 08 | 1.10E + 041.09E + 041.98E + 04 |
tries were presented. During the model development phase of ARIMA, SARIMA, RNN-GRU, and RNN-LSTM models, time-series data of each country is split into testing and training datasets. For RNN based models, the hyper parameters
during the model development phase of ARIMA, SARIMA, RNN-GRU, and RNN-LSTM models, time-series data of each country is split into testing and training datasets. For RNN based models, the hyper parameters
such as number of epochs, hidden size, number of layers, and learning rate are optimized for each country. The details of the optimized model parameters of proposed ARIMA, SARIMA, LSTM, GRU models, and their corresponding evaluation metrics (MSE and RMSE) for forecasting the cumulative confirmed cases of each country were given in Table 2. The well-trained models with optimized parameters were selected to forecast the cumulative confirmed cases until August 22nd 2021. Fig. 2 shows the 60-day forecast of the cumulative confirmed cases of the selected 10 countries as a function of time (i.e., number of days). Black line represents reported confirmed cases, the test data validation of the LSTM, GRU, ARIMA, and SARIMA were represented by red, grey, brown, and apple green color lines respectively. Similarly, the forecasts based on LSTM, GRU, ARIMA, SARIMA were depicted with dark lime green, blue, magenta, and cyan colors respectively. Even though countries such as India, the USA, UK have implemented vaccine drives and stringent preventive measures such as social distancing, lockdowns, the forecast of these countries has a steady increase. From the results, it is evident that the forecast of these 10 countries shows mostly an upward trend.

From Fig. 2, in the case of USA both statistical models (ARIMA and SARIMA), and deep learning models (LSTM and GRU) performed relatively well for the test data validation. The predictions based on ARIMA and SARIMA models did not vary much with RMSE of 50,300 and 40,700 respectively. However, the GRU and LSTM models outperformed ARIMA and SARIMA with lower RMSE of 10,951 and 12,147 respectively (Table 2). Despite of many parameters associated with the LSTM model, it required only 650 epochs to converge. However, the GRU model required 518 epochs to converge with lower MSE than that of LSTM. The USA’s optimized LSTM model has 3 hidden layers with 300 neurons per each layer whereas, optimized GRU has 2 hidden layers with 300 neurons in each layer. The 60-day forecast of confirmed cases in the USA follows a slight upward trend according to the ARIMA model, but GRU and LSTM based models predicted that the number of cases in forecast period might reach a plateau. On the other hand, the SARIMA model projected there will be a lower number of new confirmed cases when compared to the other models. These observations could be explained due to the fact that the US has a high vaccination rate. According to ARIMA (5,2,2), there will be ≈35 M confirmed cases. However, the LSTM and GRU models predicted ≈34.4 M number of cumulative confirmed cases. Whereas SARIMA forecasted that there will be ≈32.5 M of cumulative confirmed cases by the end of August 2021 in the USA which is much lower than the forecasts based ARIMA, GRU, and LSTM.

In the case of Brazil’s cumulative confirmed cases, the predictions from the ARIMA, SARIMA and GRU models are almost the same with the projection of ≈29 M. However, the LSTM model of Brazil predicted that there will be 21 M cumulative confirmed cases by the end of August 2021. LSTM and GRU models of Brazil required 1,000 epochs to converge. Moreover, other parameters - learning rate, number of hidden layers, hidden size remained the same for both GRU and LSTM models of Brazil. Interestingly, LSTM and GRU models have different values of MSE and RMSE. LSTM forecasted that the confirmed cases would approach a plateau in contrast to the ARIMA, SARIMA and GRU model’s linear increasing trend. Cumulative confirmed cases in India are on the raise to 33 M according to the SARIMA and LSTM model, whereas the ARIMA and GRU model of India has forecasted that cumulative confirmed cases would reach a total of 31 M by the end of August 2021. Both ARIMA and GRU models show that cumulative cases are approaching plateau by the last week of August 2021. For the same combination of hyper parameters, the GRU model with 894 epochs for India has lesser RMSE and MSE values when compared to that of the LSTM model with 1000 epochs. However, SARIMA(1,2,0)(1,0,7) outperformed ARIMA(1,2,1) and other models with lowest RMSE and MSE values.

The GRU and LSTM models of Russia did reasonably well however, the GRU model performed better than LSTM, with a difference in the RMSE and MSE values. The performance of the models varied based on the information embedded in the time-series data that changes from country to country. In the case of Russia, LSTM performed well with lesser RMSE (8636.29) than that of the GRU (25805.8) and the cumulative confirmed cases forecasted by LSTM were ≈30,000,000 lesser than the GRU forecast. According to the models except for LSTM, Russia will go through an exponential growth phase of the pandemic with a maximum number of cumulative confirmed cases ≈69 M by the end of August 2021. LSTM based model predicted that there will be ≈10 M lesser cases than the projections of ARIMA, SARIMA, and GRU models.

For South Africa, the number of cumulative confirmed cases are increasing at a faster pace as the trend is following an exponential growth phase since April 2021. Statistical models (ARIMA and SARIMA) predicted that the number of cases will further increase at the same pace in the 60-day forecast period. The GRU and LSTM models predicted that there will be ≈24.5 M and ≈25 M confirmed cases in South Africa respectively. However, ARIMA and SARIMA projected that cumulative cases number will be greater than ≈26 M cases with a maximum of 30 M cases by the end of August 2021.

For Mexico’s data, the ARIMA and SARIMA outperformed LSTM and GRU models with lower values of MSE and RMSE. The forecasted number of cumulative confirmed cases by ARIMA and SARIMA models of Mexico are ≈20,000,000 greater than the GRU predictions. Both LSTM and GRU models have shown a similar trend for Peru’s time-series analysis, but both ARIMA and SARIMA showed a linear upward trend regardless of the seasonality that is present in the reported cases. Deep learning model (GRU and LSTM) predictions (≈21 M) are more realistic since the forecast trend resembles the trend of reported cases in the 2021 (Fig. 2). For Chile and UK’s data, SARIMA based models performed well with RMSE values of 442 and 822 respectively. But for Iran’s data LSTM model performed better than other models with an RMSE value of 10943. For these countries, the models predicted a gradual increment in the new cases over the 60-day forecast period. SARIMA(3,2,0)(2,0,1,7) predicted that there will be ≈1,650,000 cases in Chile by the end of August 2021. Similarly, for UK’s data best model SARIMA(6,2,0) (0,2,7) predicted a maximum of ≈55,00,000. LSTM based RNN models of Chile, UK and Iran predicted that there will be ≈1750000, ≈65,00,000, ≈40,00,000 respectively. Further forecasts based on ARIMA and GRU-RNN models are presented in Fig. 2.

The results discussed above provide a clear picture of the upcoming surge of cumulative cases in the top 10 highly affected countries. Our forecast of these countries can help
| Country  | Models                        | Epochs       | Hidden size          | No. of layers | Learning rate | MSE          | RMSE         |
|---------|-------------------------------|--------------|----------------------|---------------|---------------|--------------|--------------|
| USA     | GRULSTMARIMA(4, 2, 2)SARIMA(2, 2, 3)(0, 0, 2, 7) | 2.00E + 032.00E + 03 | 3.00E + 023.00E + 02 | 2.00E + 002.00E + 00 | 1.00E-051.00E-05 | 3.30E + 102.67E + 101.63E + 106.05E + 08 | 1.82E + 051.00E + 051.28E + 04 |
| Brazil  | GRULSTMARIMA(1, 2, 2)SARIMA(3,2,3) | 1.00E + 031.00E + 03 | 3.00E + 023.00E + 02 | 2.00E + 002.00E + 00 | 1.00E-051.00E-05 | 5.71E + 095.06E + 096.21E + 094.89E + 09 | 7.56E + 047.12E + 047.88E + 046.99E + 04 |
| India   | GRULSTMARIMA(5,2,3)SARIMA(5, 2, 2) | 8.94E + 021.04E + 03 | 3.00E + 023.00E + 02 | 2.00E + 002.00E + 00 | 1.00E-051.00E-05 | 1.49E + 102.87E + 102.11E + 111.09E + 10 | 1.22E + 051.05E + 051.53E + 05 |
| Russia  | GRULSTMARIMA(5, 2, 4)SARIMA(0, 2, 1)(1, 0, 1, 7) | 1.75E + 031.19E + 03 | 3.00E + 023.00E + 02 | 2.00E + 002.00E + 00 | 1.00E-051.00E-05 | 6.31E + 087.45E + 086.68E + 079.72E + 07 | 2.51E + 042.73E + 048.17E + 039.86E + 03 |
| South Africa | GRULSTMARIMA(1, 2, 2)SARIMA(4,2,0) | 8.45E + 028.41E + 02 | 3.00E + 023.00E + 02 | 2.00E + 002.00E + 00 | 1.00E-051.00E-05 | 2.88E + 086.29E + 087.16E + 086.98E + 04 | 1.70E + 047.93E + 038.46E + 032.54E + 02 |
| Mexico  | GRULSTMARIMA(0,2,1)SARIMA(0,2,1) | 8.50E + 028.85E + 02 | 3.00E + 023.00E + 02 | 2.00E + 002.00E + 00 | 1.00E-051.00E-05 | 4.51E + 061.42E + 074.46E + 068.71E + 06 | 2.12E + 033.76E + 032.11E + 032.95E + 03 |
| Peru    | GRULSTMARIMA(1,2,1)SARIMA(5,2,1) | 1.50E + 031.40E + 03 | 3.00E + 023.00E + 02 | 2.00E + 002.00E + 00 | 1.00E-051.00E-05 | 5.24E + 061.62E + 076.50E + 066.05E + 06 | 2.29E + 034.02E + 032.55E + 032.46E + 03 |
| Chile   | GRULSTMARIMA(2,2,3)SARIMA(0, 2, 1) | 2.00E + 039.68E + 02 | 3.00E + 023.00E + 02 | 2.00E + 002.00E + 00 | 1.00E-051.00E-05 | 3.75E + 075.90E + 075.01E + 074.90E + 07 | 6.12E + 037.68E + 032.24E + 037.00E + 03 |
| UK      | GRULSTMARIMA(1, 2, 2)SARIMA(6, 2, 0)(0, 0, 2, 7) | 1.20E + 031.75E + 03 | 3.00E + 023.00E + 02 | 2.00E + 002.00E + 00 | 1.00E-051.00E-05 | 1.55E + 029.79E + 024.12E + 021.08E + 02 | 1.25E + 013.12E + 012.03E + 011.04E + 01 |
| Iran   | GRULSTMARIMA(4,2,4)SARIMA(1,2,1) | 5.18E + 026.50E + 02 | 3.00E + 023.00E + 02 | 2.00E + 002.00E + 00 | 1.00E-051.00E-05 | 1.96E + 051.94E + 052.81E + 081.99E + 08 | 4.43E + 024.41E + 021.68E + 041.36E + 04 |

Comparative analysis of Gated Recurrent Units (GRU), long Short-Term memory (LSTM) cells.
plan the healthcare policies and interpositions, moreover, the cumulative cases results provide information of the upcoming surge in the new cases, therefore, helping governments prepare for long-term and short-term responses for the pandemics such
as COVID-19. For example, according to the forecast of the cumulative confirmed cases of South Africa, there will be 30 M new cumulative confirmed cases by the end of August 2021. South Africa can avoid the upcoming surge that is projected by our models, by implementing more vaccine drives, curtailing travel between the COVID-19 hotspots and other regions within the country and limiting the social gathering at local businesses. Our forecasted results can help countries that are implementing or implemented strict vaccine drives and lockdowns. For example, India is implementing temporary lockdowns from time to time as the new surge appears in the number of cases. Based on the projected surge of the forecasted cumulative confirmed cases India can plan for opening local businesses with limited gatherings and vaccinating the majority of the population thereby decreasing the number of new cases and fatalities. The forecast of cumulative confirmed cases based on our models not only alerts these 10 countries but also help the rest of the world to prepare for the ongoing COVID-19 pandemic.

3.2. Cumulative recovered cases

Before forecasting the cumulative recovered cases in 10 of the countries that are severely affected by COVID-19 pandemic, the best ARIMA, SARIMA, GRU, and LSTM models were for each country were proposed by validation and optimization. For countries like Russia, Mexico, and Chile, ARIMA models performed better than other models that are proposed to forecast recovered cases. The RMSE of ARIMA models for these countries are ≈8170.7, ≈2112, and ≈2238 respectively. The ARIMA models of the rest of the countries the USA, Brazil, India, South Africa, Peru, the UK, and Iran have an RMSE of 127836, 78783, 372766, 84553, 2549, 20.2, and 16,753 respectively. Similarly, for recovered cases data of India, South Africa, the USA, and the UK, SARIMA based models performed better than other developed models. The RMSE of SARIMA models for these countries is ≈104581.63, ≈254, ≈24591, and ≈10.31 respectively. The SARIMA models of the rest of the countries USA, Brazil, Peru, UK, Russia, Mexico, Chile, and Iran have an RMSE of 24591, 69915, 2460.44, 10.37, 9856, 2951, 7000, and 13599.6. For the data of Brazil and Iran, the LSTM model performed better than other models with an RMSE of 71,157 and 440 respectively. For the rest of the countries such as the USA, Peru, UK, Russia, Mexico, Chile, India, South Africa with the RMSE values of the LSTM model are 163283, 4023, 31.2, 27293.02, 3762.4521, 7682, 169479, and 7930 respectively. However, for Peru’s data GRU outperformed other models with the lowest RMSE of 2289. The rest of the RMSE and MSE values of the recovered cases with respect to studied countries can be found in Table 3.

The reported recovered cases and the 60-day forecast of the recovered cases can be seen in Fig. 3. It is clear from Fig. 3, according to ARIMA based model the number of recovered cases in USA, Brazil, India, Russia, South Africa, Mexico, Peru, Chile, UK, and Iran by the end of August 2021 will be ≈12,000,000, ≈19,500,000, ≈34,000,000, ≈5,500,000, ≈2,450,000, ≈2,350,000, ≈2,750,000, ≈1,800,000, ≈1,600 and ≈3,500,000, respectively. According to SARIMA models of USA, Brazil, India, Russia, South Africa, Mexico, Peru, Chile, UK, and Iran the cumulative recovered cases will be ≈11,900,000, ≈18,500,000, ≈33,000,000, ≈5,500,000, ≈2,150,000, ≈2,100,000, ≈2,250,000, ≈1,850,000≈16,000, and ≈3,500,000, respectively. Based on the LSTM models’ predictions of the countries USA, Brazil, India, Russia, South Africa, Mexico, Peru, Chile, UK, and Iran the cumulative recovered cases will be ≈12,100,000, ≈16,990,000, ≈30,000,000, ≈5,010,000, ≈2,650,000, ≈2,650,000, ≈22,100,000, ≈1,750,000, ≈14,500, and ≈3,250,000, respectively. Similarly, based on the country-specific GRU models, the cumulative recovered cases in the countries such as USA, Brazil, India, Russia, South Africa, Mexico, Peru, Chile, UK, and Iran will be approximatly 12,120,000, ≈17,850,000, ≈30,550,000, ≈5,500,000, ≈2,850,000, ≈215,000, ≈2,110,000, ≈1,750,000, ≈15,000, and ≈3,000,000 respectively.

From Fig. 3, it is evident that there are two different possible trends that the upcoming cases can follow. The two possible trends are exponential growth in recovery cases and the other one is a plateau where there is no further increase in recovery rate. For countries USA, South Africa, Chile, and Russia all the models have shown an upward trend and similar predictions with minimum variation. For India, Brazil, UK, Iran, and Peru some of the models predicted that the cumulative recovered cases might reach a plateau, but some models predicted that these countries might continue the exponential growth phase. From Fig. 3, it is evident that the ARIMA and SARIMA models of India predicted the upcoming recoveries follows an exponential trend, but the GRU and LSTM models show a curve reaching a plateau. The stringent implementation of measures to control the spread of COVID-19 can help countries such as India reach the exponential growth phase in new recovery cases. Further, in countries such as U. K. there are fewer number of reported recovered cases (15,000) even after a year of pandemic. Also, if we examine the plots of the USA; it is evident that the reported recovered cases are not recorded after December 2020 implying that process of reporting the cumulative recovered cases are different from country to country. In USA, there are 16 states that do not report or document the recovered cases and do not have a proper definition of recovered cases. Moreover, there are 8 states in the USA where the number of hospital discharges is considered as a number of recovered cases. Some states in the USA such as South Dakota define recovered case as day-based, meaning one recovered case counted if a patient discharged from the hospital is free from any symptoms for 3–42 days. On the other hand, India has conducted a few tests per 1,000 population because of the inconsistency in the record-keeping process and definition of the recovered case, the reported cases do not represent the actual number of recovered cases.

Despite the above-mentioned discrepancies in the record-keeping and definition of recovered cases, Fig. 3 shows that the number of recovered cases is increasing in most of the countries. There are other factors along with the record-keeping process and definition of recovered cases that affect the number of recovered cases. Such factors include age, underlying health conditions, and local weather conditions. For example, the recovery percentage has decreased as the age group has increased. About 65% of the patients in the age group between 20 and 40 were recovered. As the patients age group increased to 50 and above, the percentage of recovered cases decreased to 56. The rate of transmission of the
disease and susceptibility to COVID-19 was affected by the age of the infected person. 69% of infected individuals over the age of 70 years manifested the clinical symptoms. On the other hand, only an 11% of infected individuals under 20 years of age have only 11% chance of manifestation of clinical symptoms [46].

Fig. 4 60-day ahead forecast of cumulative fatalities for top-10 countries based on RNN-GRU and RNN-LSTM models.
Table 4  Models used for forecasting cumulative death and their parameters.

| Country    | Models                        | Epochs | Hidden size | No. of layers | Learning rate | MSE         | RMSE          |
|------------|-------------------------------|--------|-------------|---------------|---------------|-------------|---------------|
| USA        | GRULSTMARIMA (5, 2, 2)        | 02     | 3.00E + 02  | 2.00E + 00    | 1.00E-05      | 3.29E + 01  | 1.81E + 01   |
| Brazil     | GRULSTMARIMA (0, 2, 2)        | 03     | 3.00E + 02  | 2.00E + 00    | 1.00E-05      | 4.26E + 01  | 2.06E + 01   |
| India      | GRULSTMARIMA (2,2)            | 02     | 3.00E + 02  | 2.00E + 00    | 1.00E-05      | 1.15E + 01  | 3.39E + 01   |
| Russia     | GRULSTMARIMA (6, 2, 4)        | 03     | 3.00E + 02  | 2.00E + 00    | 1.00E-05      | 3.60E + 01  | 6.00E + 01   |
| South Africa | GRULSTMARIMA (6, 2, 4)    | 03     | 3.00E + 02  | 2.00E + 00    | 1.00E-05      | 7.89E + 01  | 2.81E + 01   |
| Mexico     | GRULSTMARIMA (2, 1)           | 02     | 3.00E + 02  | 2.00E + 00    | 1.00E-05      | 5.00E + 01  | 7.07E + 01   |
| Peru       | GRULSTMARIMA (1,2)           | 02     | 3.00E + 02  | 2.00E + 00    | 1.00E-05      | 2.90E + 01  | 5.38E + 01   |
| Chile      | GRULSTMARIMA (5, 2, 2)        | 02     | 3.00E + 02  | 2.00E + 00    | 1.00E-05      | 1.64E + 01  | 4.05E + 01   |
| UK         | GRULSTMARIMA (6,2,4)          | 02     | 3.00E + 02  | 2.00E + 00    | 1.00E-05      | 1.33E + 01  | 3.65E + 01   |
| Iran       | GRULSTMARIMA (2,2,3)          | 02     | 3.00E + 02  | 2.00E + 00    | 1.00E-05      | 1.44E + 01  | 1.20E + 01   |
3.3. Cumulative fatalities

This section describes the forecast of cumulative fatalities in top-10 countries. Fig. 4 provides the forecasted trends of cumulative fatalities and Table 4 provides the model parameters used to build the optimized models before using them for forecasting the fatalities. For Chile’s data, SARIMA model performed better based on the evaluation metrics RMSE(53) and MSE(2860). However, Chile’s other models: ARIMA, LSTM, and GRU performed reasonably well with RMSE values of 53, 103, 405 respectively. Similarly, for the USA, Mexico, and Peru, SARIMA based models performed better than ARIMA, LSTM, and GRU. The RMSE values of the SARIMA models of the USA, Mexico, and Peru are 702, 217, and 334 respectively. The RMSE values of ARIMA(5,2,2), SARIMA(5,2,0)(1,0,1,7), LSTM and GRU of USA’s data is 407028, 702, 1490, and 1725, respectively. Whereas GRU-RNN models of countries such as India, South Africa, and Iran, performed better than other models with RMSE values of 3391, 280 and 119. The RMSE values of the GRU models of the rest of the countries Brazil, Russia, Mexico, Peru, Chile, and the UK are 2063, 599, 707, 5382, 405 and 36 respectively. On the other hand, LSTM-based models outperformed other models for Brazil, Russia, and UK data as seen from Table 4.

According to Fig. 4, the number of cumulative fatalities in the USA will be between 600,000 and 640,000 according to forecasts of the ARIMA(5,2,2), SARIMA(5,2,0)(1,0,1,7), LSTM and GRU models. Based on the developed models, the fatalities in all the countries are increasing as the number of days into the pandemic is increasing. However, from the Fig. 4, it is evident from the trend of the forecast, that the fatalities in the USA, Peru, UK, and Mexico might reach a plateau. According to the forecasted cumulative fatalities (Fig. 4) based on GRU model with lowest RMSE of the countries - India, South Africa, and Iran, there will be $\approx550,000$, $\approx70,000$, $\approx85,000$, respectively. Whereas the statistical models based on seasonal ARIMA or SARIMA model of these countries predicted that there will be $\approx590,000$, $\approx75,000$, $\approx88,000$ cumulative death cases by the end of August 2021. Similarly, respective ARIMA models of these countries’ forecasts can be seen in Fig. 4. Whereas according to LSTM based models, fatalities in other countries – USA, Brazil, Russia, Mexico, Peru, Chile and UK, are $\approx610,000$, $\approx590,000$, $\approx169,000$, $\approx223,000$, $\approx37,000$, $\approx124,000$ respectively. The corresponding RMSE values of the LSTM models of the countries are described in Table 4. The number of cumulative fatalities forecasted by the ARIMA based models of the USA, Brazil, Russia, Mexico, Peru, Chile, and the UK, are approximately a maximum of 625,000, $\approx650,000$, $\approx160,650$, $\approx70,000$, $\approx220,000$, $\approx190,000$, $\approx38,000$ and $\approx125,000$, respectively. Similarly, the SARIMA based models’ forecasts of these countries can be seen in Fig. 4. The forecasted trends for UK, Peru, and the USA are almost the same for both statistical models (ARIMA and SARIMA) and deep learning models (LSTM and GRU) and all the models show that the cumulative fatalities will reach a plateau by the end of August 2021. Fatalities in the rest of the countries varied from model to model based on the trend of the reported data.

ARIMA of Brazil, India, and Peru show an exponential growth for the next 60 days. GRU models captured seasonality well when compared to the SARIMA models of the reported cases and forecasted well for most of the countries. For example, the forecast of Chile, Brazil, and India based on GRU-RNN showed clear seasonality in the upward trend of the forecast. Forecasts based on models of South Africa, India, Russia, and Brazil varied much when compared to the models of other countries. For example, according to ARIMA and SARIMA of models for India, the number of cumulative death cases reaches a plateau, but this can only happen if and only if strict COVID-19 prevention and vaccination policies are implemented. Upcoming new number of fatalities can be prevented by providing better health care facilities: increasing the availability of infrastructure such as the number of Intensive Care Units (ICU), number of hospital beds [47], available healthcare workers per number of patients [48] and by implementing strict social distancing, and other preventive measures in countries such as South Africa, Mexico, Peru, Brazil, India. COVID-19 can be fatal to senior citizens of any nation. Centers for Disease Control and Prevention (CDC) has described that older people are at higher risk of hospitalization due to COVID-19 and it can be fatal to them when compared to the younger population [49]. For example, there are 54.3 million residents who are 65 years and older than that [50] to whom COVID-19 can cause severe illness and can be fatal when compared to younger people of the population. Social distancing and other preventive measures have direct effect on the rise of the number of cumulative deaths which is evident from the reported data of India until April 2021 (Fig. 2). In May 2021, the number of cumulative confirmed cases and death cases (Fig. 4) increased significantly from the first reported case in India. Thus, emphasizing the importance of preventive measures and other factors that contribute to the increasing number of cumulative COVID-19 cases.

4. Conclusions

Time-series data of COVID-19 is found to be highly dynamic and the embedded information in the reported cases changes with time and country. The COVID-19 data of some countries is non-linear, for instance, reported deaths of all countries except Chile, we can observe a linear exponential relationship. ARIMA models were found to be suitable and performed well in modeling data that followed a linear relationship. For the confirmed cases of Chile and Brazil and for the recovered cases of the USA and Iran, the ARIMA model performed well. While RNN based models performed better with countries that have a non-linear relationship. Countries such as India, South Africa, and the UK have the reported cases with a non-linear relationship where deep learning models GRU-RNN, GRU-RNN and LSTM models (Table 4) performed better than SARIMA and ARIMA models respectively.

For confirmed cases, SARIMA based models performed better for India, Russia, Peru, Chile, and the UK and ARIMA model performed better for Brazil. For Mexico and Iran, LSTM model performed better and the GRU model performed better for the USA and South Africa. According to the LSTM and GRU models, the confirmed cases in the USA might reach a plateau. However, the cases are going to increase in USA at an alarming rate according to ARIMA model, but SARIMA model predicted the new number of confirmed cases might decrease. Similarly, the GRU, ARIMA, and SARIMA predicted that the confirmed cases in Brazil will
be increasing at a higher rate, but the LSTM model predicted that the new number of cases will decrease. For India, the cumulative confirmed either increase or reach plateau according to the models developed. For countries such as South Africa, Mexico, and Iran, GRU model predicted that the cumulative cases would reach a plateau but the rest of the models (LSTM, ARIMA, and SARIMA) predicted that the number of new cases in these countries will increase. LSTM model predicted that the cases in Russia will decrease but models based on ARIMA, SARIMA and GRU predicted that the cumulative cases in Russia are increasing at a faster pace instead. For Peru, the ARIMA and SARIMA models predicted an exponential growth in cumulative confirmed cases. However, LSTM and GRU models predicted that cumulative cases might reach a plateau in the same period. For Chile’s confirmed cases, the LSTM and ARIMA models forecasted that the cases would alarmingly increase and models SARIMA and GRU predicted that cumulative confirmed cases in Chile would reach a plateau. However, for UK’s confirmed cases, all models predicted that cumulative cases are going to increase at a faster pace.

For recovered cases, the ARIMA model performs better for the USA, Russia, and Chile. Whereas the SARIMA model performed better for Brazil, India, and South Africa. For Mexico, Peru and the UK, the GRU model performed better than other models. For Iran, the LSTM model performed better than other models. The recovered cases in USA, South Africa, and Chile will increase according to all the models. However, the GRU and LSTM models predicted that recovered cases in Peru and UK will reach a plateau. For Iran, except for GRU model, all other models predicted exponential growth in the recovered cases. For Brazil and India, the ARIMA and SARIMA model predicted an exponential increase in recovered cases for the next 60 days. Whereas GRU and LSTM predicted the forecast might reach a plateau. For Russia, all models except LSTM predicted a continuous increase in recovered cases. Whereas with Iran’s recovered data, except the GRU model, all other models predicted that the recovery rate in the country will increase in the next 60 days.

For the forecast of fatalities data, the LSTM model outperformed the rest of the developed models for Brazil, Russia, South Africa, Peru, and the UK. For countries India and Iran, GRU models outperformed the rest of the developed models. On the other hand, SARIMA based models performed well for the USA, Mexico, and Chile. ARIMA models of all countries have higher RMSE and MSE values indicating that ARIMA models are not suitable for the COVID-19 fatalities. The reported fatalities have a non-linear relationship that cannot be addressed by simple ARIMA models. Therefore, SARIMA models that can capture seasonality and RNN based models performed better with countries that have a non-linear relationship. Shockingly, the fatalities in South Africa, Russia, Chile, and Iran will be continuing to increase with respect to all models. Similarly, the recovered cases might reach a plateau in next 60 days according to all models of countries like UK, Mexico (except ARIMA), India (except ARIMA and SARIMA) and USA (except ARIMA). Based on the results and interpretations it is highly recommended to develop ARIMA models as the initial models to forecast the time-series data because of their simplicity, if the residual plots contain non-linear relationships, then more complex models such as RNN-LSTM and RNN-GRU can be developed. More amount of data with high accuracy is needed to develop accurate and robust models to obtain forecasts with less margin of error and for correlating the forecasts with the factors that contribute to COVID-19’s rapid spread. We conclude that it is always recommended to build simple statistical models such as ARIMA and SARIMA before developing more complex models such as LSTM and GRU.

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.

References

[1] D. Cucinotta, M. Vanelli, WHO declares COVID-19 a pandemic, Acta Bio-med.: Atenei Parmensis 91 (1) (2020) 157–160. https://doi.org/10.23750/abm.v91i1.9397.
[2] Worldometers.info, 2020, https://www.worldometers.info/.
[3] H. Swapnarekha, H.S. Behera, J. Nayak, B. Naik, Role of intelligent computing in COVID-19 prognosis: a state-of-the-art review, Chaos, Solitons Fractals 138 (2020) 109947, https://doi.org/10.1016/j.chaos.2020.109947.
[4] S. Ghosal et al, Prediction of the number of deaths in India due to SARS-CoV-2 at 5–6 weeks, Diab. Metab. Synd.: Clin. Res. Rev. (2020), https://doi.org/10.1016/j.dsx.2020.03.017.
[5] D. Parbat, M. Chakraborty, A Python based Support Vector Regression Model for prediction of Covid19 cases in India, Chaos, Solitons Fractals 138 (2020) 109942, https://doi.org/10.1016/j.chaos.2020.109942.
[6] M. Maleki, M.R. Mokhtari, D. Wrath, K.-H. Pho, Time series modelling to forecast the confirmed and recovered cases of COVID-19, Travel Med. Infect. Dis. 37 (2020) 101742, https://doi.org/10.1016/j.tmtd.2020.101742.
[7] D. Benvenuto, M. Giovanetti, L. Vassallo, S. Angeletti, M. Ciccozzi, Application of the ARIMA model on the COVID-2019 epidemic dataset, Data Brief 29 (2020) 105340, https://doi.org/10.1016/j.dib.2020.105340.
[8] M.H.D.M. Ribeiro, R.G. da Silva, V.C. Mariani, L.D.S. Coelho, Short-term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil, Chaos, Solitons Fractals 135 (2020) 109853, https://doi.org/10.1016/j.chaos.2020.109853.
[9] P. Kumar et al, Forecasting the dynamics of COVID-19 Pandemic in Top 15 countries in April 2020: ARIMA Model with Machine Learning Approach, medRxiv (2020), https://doi.org/10.1101/2020.03.30.20046227.
[10] S.F. Ardabili, et al., Covid-19 outbreak prediction with machine learning. Available at SSRN 3580188, 2020. DOI: https://doi.org/10.3390/ai13100249.
[11] V.K.R. Chimmula, L. Zhang, Time series forecasting of covid-19 transmission in Canada using LSTM networks, Chaos, Solitons Fractals 135 (2020) 109864, https://doi.org/10.1016/j.chaos.2020.109864.
[12] R. Salgotra, M. Gandomi, A.H. Gandomi, Time series analysis and forecast of the COVID-19 pandemic in India using genetic programming, Chaos, Solitons Fractals 138 (2020) 109945, https://doi.org/10.1016/j.chaos.2020.109945.
[13] H. Qi, S. Xiao, R. Shi, M.P. Ward, Y. Chen, W. Tu, Q. Su, W. Wang, X. Wang, Z. Zhang, COVID-19 transmission in Mainland China is associated with temperature and humidity: A time-series analysis, Sci. Total Environ. 728 (2020) 138778, https://doi.org/10.1016/j.scitotenv.2020.138778.
[14] I. Kırbaş et al, Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN...
and LSTM approaches, Chaos, Solitons Fractals (2020), https://doi.org/10.1016/j.chaos.2020.110015 110015.

[15] S.K. Bandyopadhyay, S. Dutta, Machine learning approach for confirmation of covid-19 cases: positive, negative, death and release, medRxiv (2020), https://doi.org/10.1101/2020.03.25.20043505.

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

[17] M. Azarafza, M. Azarafza, J. Tanha, COVID-19 infection forecasting based on deep learning in Iran, medRxiv (2020), https://doi.org/10.1101/2020.05.16.20104182.

[18] G. Perone, Comparison of ARIMA, ETS, NNAR, TBATS and hybrid models to forecast the second wave of COVID-19 hospitalizations in Italy, Eur. J. Health Econ. (2021) 1–24, https://doi.org/10.1007/s10198-021-01347-4.

[19] M. Ala'raj, M. Majdalawieh, N. Nizamuddin, Modeling and forecasting of COVID-19 using a hybrid dynamic model based on SEIRD with ARIMA corrections, Infect. Dis. Model. 6 (2021) 96–111, https://doi.org/10.1016/j.idm.2020.11.007.

[20] O. Orojeh, T. Shuggish State-Based Neural Networks Provide State-of-the-art Forecasts of Covid-19 Cases, International Conference on Applied Intelligence and Informatics, Springer, 2021.

[21] J. s. r, r., et al, ARIMA based Time Series Analysis: Forecasting COVID-19 Most Vaccinated Process and Active Cases classify using Probability Distribution Curve Rates (ARIMAPDC), in: 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), 2021, https://doi.org/10.1109/ICOSEC51865.2021.9591774.

[22] J. Sun, Forecasting COVID-19 pandemic in Alberta, Canada using modified ARIMA models, Comput. Meth. Prog. Biomed. Update 1 (2021) 100029, https://doi.org/10.1016/j.cmpbup.2021.100029.

[23] S. Roy, G.S. Bhunia, P.K. Shit, Spatial prediction of COVID-19 epidemic using ARIMA techniques in India, Model. Earth Syst. Environ. 7 (2) (2021) 1385–1391, https://doi.org/10.1007/s40808-020-00890-y.

[24] J. Hopkins, Novel Coronavirus (COVID-19) Cases Data. 2020 [cited 2020; Available from: https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases.

[25] H. Tandon, et al, Coronavirus (COVID-19): ARIMA based cumulative COVID-19 cases (confirmed, recovered and deaths) forecasting: a deep learning approach, Int. J. Forecast. 35 (4) (2019) 1548–1560, https://doi.org/10.1016/j.ijforecast.2018.07.006.

[26] S. Abolmaali, S. Shirzaei, Forecasting COVID-19 number of hospitalizations in Italy, Eur. J. Health Econ. (2021) 1–24, https://doi.org/10.1038/s41598-021-97037-5.

[27] K.E. ArunKumar, D.V. Kalaga, C.M. Sai Kumar, G. Chilkoor, K.N. Nabi, M.T. Tahmid, A. Rafi, M.E. Kader, M.A. Haider, Forecasting COVID-19 cases: a comparative analysis between recurrent and convolutional neural networks, Results Phys. 24 (2021) 104137, https://doi.org/10.1016/j.rinp.2021.104137.

[28] P. Kang-Lin, C.-H. Wu, J.G. Yeong-Jia, The development of a new statistical technique for relating financial information to stock market returns, Int. J. Manage. 21 (4) (2004) 492, https://doi.org/10.1186/s40854-019-0157-x.

[29] M.P.S. Chawla, H.K. Verma, V. Kumar, RETRACTED: a new statistical PCA–ICA algorithm for location of R-peaks in ECG, Int. J. Cardiol. 129 (1) (2008) 146–148.

[30] X. Li, W. Shang, S. Wang, Text-based crude oil price forecasting: a deep learning approach, Int. J. Forecast. 35 (4) (2019) 1548–1560, https://doi.org/10.1016/j.ijforecast.2018.07.006.

[31] J. Chung, et al., Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.5555, 2014.

[32] K.E. ArunKumar, D.V. Kalaga, C.M. Sai Kumar, G. Chilkoor, M. Kawaji, T.M. Brenza, Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA), Appl. Soft Comput. 103 (2021) 107161, https://doi.org/10.1016/j.asoc.2021.107161.

[33] pmidarima/8.2, april 19 2021; Available from: https://ppypi.org/project/pmidarima/.

[34] Z. Malki, E.-S. Atlam, A. Ewis, G. Dagne, A.R. Alzighaibi, G. Elnahromy, M.A. Elhosseini, A.E. Hassani, I. Gad, ARIMA models for predicting the end of COVID-19 pandemic and the risk of second rebound, Neural Comput. Appl. 33 (7) (2021) 2929–2948, https://doi.org/10.1007/s00521-020-05434-0.

[35] Dylan Tiger, et al., Reports on “recovered” Covid-19 cases inconsistent and incomplete, Numbers elusive and may mislead on real medical impact of virus. Informatics in Medicine Unlocked 2020 [cited 2020 10 july]; Available from: https://www.eu-citizenaccess.org/2020/07/10/definitions-of-recovered-from-covid-19-vary-widely-across-the-u-s/.

[36] R. Tyagi, L.K. Dwivedi, A. Sanzgiri, Estimation of Effective Reproduction Numbers for COVID-19 using Real-Time Bayesian Method for India and its States, 2020, Paper.

[37] I. Voinsky, G. Baristaite, D. Gurwitz, Effects of age and sex on recovery from COVID-19: analysis of 5769 Israeli patients, J. Infect. 81 (2) (2020) e102–e103, https://doi.org/10.21203/rs.3.rs-45937/v1.

[38] N.G. Davies et al, Age-dependent effects in the transmission and control of COVID-19 epidemics, Nat. Med. 26 (8) (2020) 1205–1211, https://doi.org/10.1038/s41591-020-0962-9.

[39] R.K. Wadhwa, P. Wadhwa, P. Gupta, P.F. Figueroa, K.E. Joynt Maddox, R.W. Yeh, C. Shen, Variation in COVID-19 hospitalizations and deaths across New York City boroughs, Jama 323 (21) (2020) 2192, https://doi.org/10.1001/jama.2020.7197.
[48] Organization, W.H., *Coronavirus disease 2019 (COVID-19): situation report, 82.* 2020.

[49] CDC. *Older Adults.* 2020 [cited 2020 11 SEPTEMBER ]; Available from: https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/older-adults.html.

[50] *Older Population and Aging.* 2019 [cited 2019; Available from: https://www.census.gov/topics/population/older-aging.html].