International Journal of Current Research and Review
DOI: http://dx.doi.org/10.31782/IJCRR.2021.SP191

Scopus

Research Article

IJCRR
Section: Healthcare
ISI Impact Factor (2019-20): 1.628
IC Value (2019): 90.81
SJIF (2020) = 7.893

Covid-19 Forecasting and Analysis Using Different Time - Series Model and Algorithms

Yash Srivastava, Shikhar Bhardwaj, Parvathi R

School of Computing Science and Engineering Vellore Institute of Technology, Chennai, India.

ABSTRACT

Introduction: The novel coronavirus (COVID-19) has significantly spread over the world and impacted with new challenges to the research community. Although governments initiated numerous containment and social distancing measures all over the world, the need for healthcare resources has dramatically increased and the effective management of infected patients becomes a challenging task for healthcare centres.

Objective: Thus, the objective of the research is to find the accurate short-term forecasting of the number of new confirmed covid-19 positive cases is important for optimizing the available resources and slowing down the progression of COVID-19. Recently, various methods like machine learning models and other algorithms demonstrated important improvements when handling time-series data in various applications.

Methods: This paper presents a comparative study of different machine learning methods and models to forecast the number of new cases. Specifically, Long short-term memory (LSTM), Autoregressive Integrated Moving Average (ARIMA), Holt's Linear forecasting model, Exponential smoothing and Moving-average model algorithms have been applied for forecasting of COVID-19 cases based on data set.

Result: Results were analysed using various parameters like Root Mean Square Error, Mean Absolute Error, Mean Absolute Percentage Error, Error Vector Magnitude Root Mean Square Logarithmic Error.

Conclusion: As a conclusion, compared to other models, Long Short TermModel predicted better forecasting and gives the best performance in terms of different parameters.

Key Words: Data-driven, Deep learning, COVID-19 Forecasting, Long short-term memory, Moving average, Exponential smoothing, Holts forecasting model

INTRODUCTION

At the end of 2019, a new coronavirus called Corona-virus Disease 2019 (COVID-19) has appeared in Wuhan city in China. Recently, the COVID-1 is flagged as a pandemic by the World Health Organization on March 11th, after overpassing 118,0 0 0 cases in over 110 countries at that time. This virus has exponentially spread all over the world and deeply affected healthcare systems in many countries, such as Italy, Spain, France, and the United States, India. The increased demand for healthcare resources generated a large number of patients lead to hospital resources shortages and intensified situations in hospitals. Accurately modelling and forecasting the spread of confirmed COVID-19 new cases is vital to understand and help decision-makers to slow down its impact of spreading. Currently COVID-19 pandemic is one of the most serious diseases confronting our world because of its highly negative effects on health over a large number of people.[1] Its impact is noticeable on sensitive populations, including the aged and peoples with chronic health conditions, such as asthmatics. Hence, it becomes a multidisciplinary issue that involves both the epidemiological experts, pharmaceutical industry, specialists in modelling diagnosis systems. There have been many modelling approaches presented by researchers for different regions like India[1], Pakistan, USA, China, Italy and other countries. Different studies have been carried out using time series forecasting models like ARIMA[2-4] and Exponential

Corresponding Author:

Yash Srivastava, School of Computing Science and Engineering Vellore Institute of Technology, Chennai, India.

ISSN: 2231-2196 (Print) **ISSN:** 0975-5241 (Online)

Received: 31.12.2020 **Revised:** 19.01.2021 **Accepted:** 11.02.2021 **Published:** 30.03.2021

Smoothing. This paper is within the field of modelling and forecasting of COVID-19 time-series data in which different approaches for forecasting the time series of upcoming dates is carried out and compared for different algorithms such as Long short-term memory (LSTM).[5] Autoregressive Integrated Moving Average (ARIMA), Holt's Linear forecasting model.[7] Exponential smoothing and Moving-average model algorithms. All of the mentioned algorithms are trained and tested on the covid-19 dataset collected from Kaggle 2020 and CORD. The collected data is preprocessed and modelled before applying it to the different algorithms to get the predictions more accurately. It is very useful to forecast the number of Covid-19 cases so that resources in the hospital can be optimally managed, which in turn saves the many infected patients. Recently, machine learning and deep learning have emerged as a promising field of research in a wide range of applications, both in academia and industry.[7,8]

The first 100 days: Modeling the evolution of the COVID-19 pandemic

This paper features an analytical model for modelling the evolution of the 2020 COVID-19 pandemic. The model featured is based on the numerical solution of the widely used Susceptible-Infectious-Removed (SIR) populations model for describing epidemics. An expanded version of the original Kermack McKendrick Model is considered, which includes a decaying value of the parameter (the effective contact rate), interpreted as an effect of externally imposed conditions, to which it referred as the forced-SIR (FSIR) model. The proposed model contains 3 adjustable parameters which are obtained by fitting actual data (up to April 28, 2020). The results are analysed results to infer the physical meaning of the parameters involved. The model relies on only three parameters all of which are obtained by directly fitting the reported data of daily populations of infected individuals. The model is used to make predictions about the total expected number of infections in each country as well as the date when the number of infections will have reached 99% of this total. We also compare key findings of the model with recently reported results on the high contagiousness and rapid spread of the disease.[9,10]

As a model of simple SIR type, it does not take into consideration age, gender, spatial position or any other factors. It assumes homogeneous mixing, that is, individuals make contact at random, the transmission and recovery rates are positive and the same for all individuals, there is no vaccine available, the total population size is constant and large, and any recovered person obtains permanent immunity. It can also be estimated as to how each country has been affected, describing the severity of the situation. The model has certain limitations, such as not being able to make any predictions about the mortality rate, and restricting the compartmentalization of the population in only 3 classes (S, I, R).

Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study

This paper depicts the analysis of various deep learning methods to predict the number of new cases and recovered cases of Covid-19. Uni-variate time series data of daily confirmed and recovered cases from 6 countries Italy, Spain, France, the USA, China, and Australia was the sector of focus. Each model was first trained with training measurements and then each variable was predicted using the training models for the unseen testing dataset. The training data consist of univariate time series data of confirmed and recovered cases from January 22, 2020, through May 31, 2020. It was found that RNN is relatively faster than the other models followed by GRU. This is mainly because the RNN is a simple model and the GRU use directly all hidden states without control and presents fewer computational parameters compared to LSTM, Bi-LSTM and VAE. The method VAE provides better forecasting of COVID-19 confirmed cases in comparison to the other considered models for almost all considered countries. The VAE model outperformed the other models by providing good forecasting performance with lower RMSE, MAE, MAPE and RMSLE, and EV values. This fact is maybe due to the capacity of the VAE in dealing with small data compared to the other recurrent models (RNN, LSTM, Bi-LSTM, and GRU) which may need more lengthy data to extract relevant variability in time series data. On the other hand, RNN and its improved versions LSTM, BiLSTM, and GRU provide relatively moderate forecasting performance in terms of the evaluation metrics (RMSE, RMSE, MAPE, and RMSEL) and perform very poorly in terms of explained variance. This may be explained by the lack of a good amount of training data needed to capture the COVID-19 data dynamics. The VAE can capture almost all variability in data and provide more accurate forecasting in comparison to the other RNN-based models. All other models perform moderate forecasting performance in terms of RMSE, MAE, MAPE, and RMSLE and show poor performance in terms of explained variance. This is maybe due to their need for more data in the training to capture the dynamics of COVID-19. The worst model is RNN because of its simplicity and followed by its extended versions of LSTM, Bi-LSTM, and GRU models.[11,12]

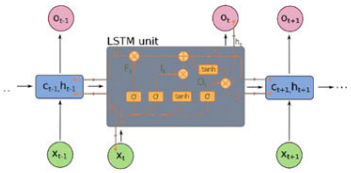
MATERIALS AND METHOS

Algorithms and Models

1) Long short-term memory (LSTM)

Long Short_Term Memory is a kind of Artificial Recurrent Neural Network (RNN) architecture and it has a feedback connection. The gate structure of the LSTM is given in Equation 1. It will process the images, speech and video. Example task of LSTM is unsegmented, handwriting rec-

ognition, speech recognition, traffic detection and Intrusion detection systems.[4-8]



$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i)$$
$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f)$$
$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o)$$

Eq. 1: LSTM gates.

$i_t \rightarrow$ *represents input gate.*
$f_t \rightarrow$ *represents forget gate.*
$o_t \rightarrow$ *represents output gate.*
$\sigma \rightarrow$ *represents sigmoid function.*
$w_x \rightarrow$ *weight for the respective gate(x) neurons.*
$h_{t-1} \rightarrow$ *output of the previous lstm block(at timestamp t − 1).*
$x_t \rightarrow$ *input at current timestamp.*
$b_x \rightarrow$ *biases for the respective gates(x).*

2) Autoregressive Integrated Moving Average (ARIMA) model

Box and Jenkins popularized a method that combines both autoregressive (AR) and moving average (MA) models. An ARMA (p,q) model is a combination of AR(p) and MA(q) models and is best used for univariate time series modelling. In the AR(p) model, the future value of a variable is assumed to be dependent upon a linear combination of p past observations and a random error term.

Autoregressive integrated moving average (ARIMA) models were popularized by Box and Jenkins (1970 ARIMA model is a mixture of autoregression (AR) and Moving Average (MA) models which use the autocorrelation of the time series data. It is also used to find the changes that happened in the stationary data. ARIMA(p,d,q) is a times series model function that contains p as autoregression, d is the degree of difference and q is the number of moving average represented in Equation 2.[13,14]

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_0 Y_0 + \epsilon_t$$
$$Y_{t-1} = \beta_1 Y_{t-2} + \beta_2 Y_{t-3} + .. + \beta_0 Y_0 + \epsilon_{t-1}$$

Eq. 2: **ARI**MA model.

ARIMA model follows a Box-Jenkins approach where changes in the stationary models are accessed using information criteria and autocorrelation. It also supports the latent variable estimates and autocorrelation residuals. Once it's good, it can be used for retrospection and forecasting.[15]

3) Exponential smoothing

The exponential smoothing method uses a different type of "smoothing" which differs from average smoothing. The previous time steps are exponentially weighted and added up to generate the forecast. The weights decay as we move further backwards in time. The model can be summarized in Equation 3 as follows:

$$\hat{y}_{t+1} = \alpha \cdot y_t + \alpha \cdot (1-\alpha) \cdot y_{t-1} + \alpha \cdot (1-\alpha)^2 \cdot y_{t-2} + ...$$
$$\hat{y}_{t+1} = \alpha \cdot y_t + (1-\alpha) \cdot \hat{y}_t$$

Eq. 3: Exponential smoothing.

In the above equations, α is the smoothing parameter. The forecast yt+1 is a weighted average of all the observations in the series y1, …, it. The rate at which the weights decay is controlled by the parameter α. This method gives different weightage to different time steps, instead of giving the same weightage to all-time steps (like the moving average method). This ensures that recent cases data is given more importance than old cases data while making the forecast.[16]

4) Holt's linear model

Holt's linear is completely different from the first two methods. Holt's linear attempts to capture the high-level trends in the time series data and fits the data with a straight line. The method can be summarized in Equation 4.Forecast, level, and trend equations respectively

$$\hat{y}_{t+h} = l_t + h \cdot b_t$$
$$l_t = a \cdot y_t + (1-\alpha) \cdot (l_{t-1} + b_{t-1})$$
$$b_t = \beta \cdot (l_t - l_{t-1}) + (1-\beta) \cdot b_{t-1}$$

Eq. 4: Holt's Linear.

In the above equations, αand β are constants that can be configured. The values lt and bt represent the level and trend values respectively. The trend value is the slope of the linear forecast function and the level value is the y-intercept of the linear forecast function. The slope and y-intercept values are continuously updated using the second and third update equations. Finally, the slope and y-intercept are used to calculate the forecast yt+h (in Equation 1), which is h time steps ahead of the current time step.[17]

5) Moving average Algorithm

The moving average method is more complex than the naive approach. It calculates the mean sales over the previous 30 days and forecasts that as the next day's sales. This method considers the previous 30 timesteps, and is, therefore, less prone to short term fluctuations than the naive approach. The

model can be summarized in Equation 5:

$$\hat{y}_{t+1} = \frac{1}{30} \cdot \sum_{t-30}^{t} y_n$$

Eq. 5: Moving average.

In the above equation, yt+1 is tomorrow's sales. On the right-hand side, all the sales for the previous 30 days are added up and divided by 30 to find the average. This forms the model's prediction, yt+1.[18]

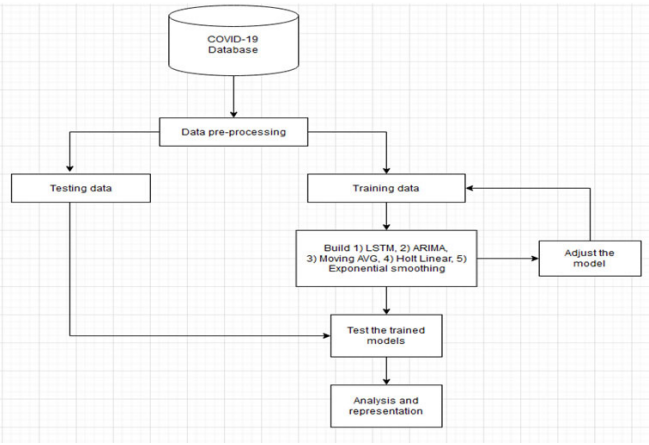


Figure 1: Conceptual framework of the proposed forecasting methods.

The COVID-19 forecasting has been done in two main stages as shown in Figure 1. training and testing. In the first stage, the raw data is preprocessed and standardized and then it is used to construct the deep learning model. The values of parameters of machine learning models are selected such that the loss function is minimized during the training. Here, Adam optimizer is used for this purpose. After that, in the testing stage, the previously constructed models with the selected parameters are used to forecast the number of COVID-19 cases. Then a proper forecasting and graphical representation are proposed as shown in Figure 1 and the model is evaluated. The accuracy of the model will be verified by comparing the measured data with real data via different statistical indicators.[18]

RESULTS AND DISCUSSION

Data Description

The COVID-19 disease was reported by the WHO in around 210 countries and territories worldwide.[5] In particular, many countries of Europe and North America suffer from a large COVID-19 outbreak. The role of large air traffic between Asia, North America, and Europe has significantly facilitated the propagation of COVID- 19 from its origin to the other infected countries; person-to-person spread was subsequently reported among close contacts of returned travellers. The total number of confirmed cases, deaths in India concerning dates are given the Figure 2. Covid-19 open Research Dataset created by the White House and various research groups contains 200,000 articles with 100,000 full texts describing the COVID-19, SARS-CoV-2, and coronaviruses. Various Analysis can be done on this open-source dataset with help of Natural Language Processing and Artificial Intelligence Techniques. These techniques are the evergreen methods used in the medical field.

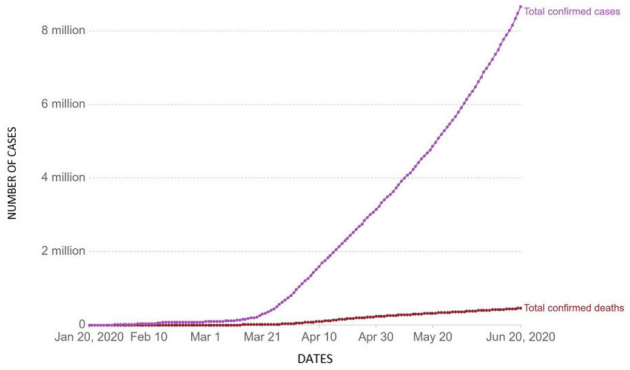


Figure 2: Covid-19 Dataset representation.

Data analysis and modelling

The gathered data is preprocessed and the daily new cases column is being separated from others and given as an input to the different models with dates included every model is tested and results are produced for comparison such as Visible units, Latent dimensions, Learning rate, Training epochs etc. Figure 3 displays the evolutions of the loss function as a function of the number of epochs in LSTM, ARIMA, Moving Average, Holt's Linear model and Exponential Smoothing during the training stage. It can be seen that the three models (LSTM and ARIMA) converge very quickly and the Moving Average model is comparatively faster than the other models. This is mainly because there are less internal and external computations in the Moving average Model.[11-14]

Forecasting Results

Each model is trained with the training measurements. Then, forecast each variable using the trained models for the unseen testing dataset. The training data consist of univariate time series data of confirmed and recovered cases from January 22, 2020, through May 31, 2020. Here, the challenge in this study is to investigate the performance of these five models in the presence of relatively small data as shown in Figure 4. Parameters and settings of the constructed LSTM, ARIMA, MA, ES and HL models based on training datasets are presented in Table 1.[13,14]

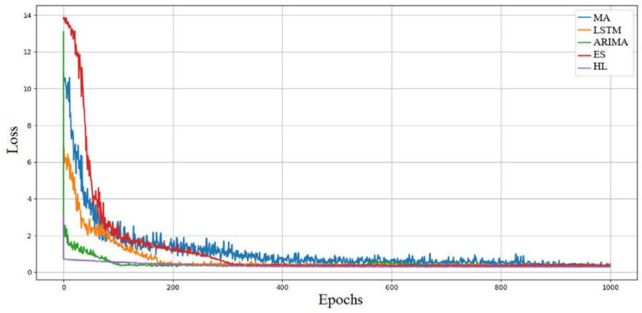


Figure 3: Convergence of the loss function of Moving Average (MA), Long short-term memory (LSTM), Autoregressive Integrated Moving Average (ARIMA), Exponential Smoothing (ES), Holt's Linear (HL) models during the training stage.

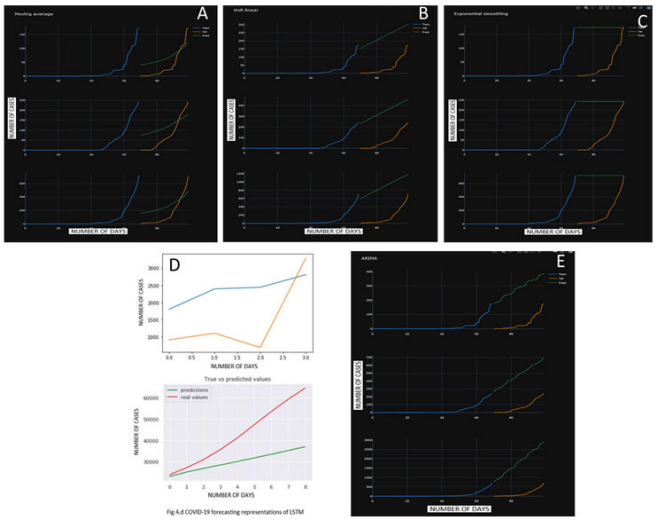


Figure 4: A. Covid-19 forecasting representation of moving average, B. Covid-19 forecasting representation of Hold Linear, C. Covid-19 forecasting representation of Exponential smooting, D. COVID-19 forecasting representations of LSTM and E. COVID-19 forecasting representations of ARIMA.

Table 1: Parameters settings of studied approaches

Models	Parameters	Value
LSTM	Learning rate	0.0005
	Timestep	05
	Training epochs	16
	Hidden units	1000
	Layers	02
ARIMA	Learning rate	0.0005
	Timestep	05
	Features	01
	Training epochs	1000
	Hidden units	16
Moving Average	Learning rate	0.0005
	Timestep	05
	Features	01
	Training epochs	1000
	Hidden units	16
Exponential Smoothing	Learning rate	0.0005
	Timestep	05
	Features	01
	Training epochs	1000
	Hidden units	16
Holt's Linear	Learning rate	0.0005
	Timestep	05
	Features	01
	Training epochs	1000
	Hidden units	16

Now, the forecasting quality of the previously designed models will be verified using unseen testing data. The testing data consists of confirmed and recovered COVID-19 new cases recorded from 1st June to 7th June 2020. After testing each mode's different validation metrics such as RMSE, MAE, MAPE, EV and RMSLE values are calculated presented in Table 2.

Table 2: Validation Metrics for confirmed cases COVID19 forecasting using LSTM, ARIMA, Moving Average, Holt's Linear and Exponential Smoothing models

Model	RMSE	MAE	MAPE	EV	RMSLE
LSTM	2,103,966	2,150,293	13,337	0779	00164
ARIMA	3,002,053	3,021,335	22,049	0002	00701
Moving Average	3,892,321	3,882,867	22,689	0017	00707
Exponential Smoothing	4,001,709	3,950,391	23,738	0103	00752
Holt's Linear	3,887,026	3,821,325	22,849	0002	00703

From Table 2 it can be seen that the LSTM model performed best among other considered models by providing better forecasting performance with lower RMSE, MAE, MAPE, RMLSE and EV values closer to 1 representing most of the variance in the data. The LSTM can catch almost all variability in data and provide more accurate forecasting in comparison to the other models considered for this study. All other models perform moderate forecasting performance in terms of RMSE, MAE, MAPE, and RMSLE and showed bad performance in terms of explained variance. The cause can be their need for more data in the training to collect the variability and dynamics of COVID-19. The worst model is Exponential Smoothing because of its simplicity and less accurate distinctions.[12-14]

CONCLUSION

The COVID-19 pandemic is exponentially spreading still in the world, and impacted the healthcare reach due to over increase in the number of patients in the hospital hence forecasting the number of new confirmed cases might help in preparing the hospitals to get facilities to accommodate and get resources for incoming patients. Overall this study provided a comparison between different approaches and models to forecast the time series of COVID-19 new confirmed cases each day. In this study different machine learning and deep learning models have been applied such as LSTM, ARIMA, Moving average, Exponential smoothing and Holt's Linear model applied to the real-time dataset of the daily COVID-19 confirmed cases. Seven days-ahead forecasts are provided based on historical data of 148 days since January 22, 2020, for the world. The performance of each model has been evaluated in terms of Root Mean Square Error, Mean Absolute Error, Mean Absolute Percentage Error, Error Vector Magnitude Root Mean Square Logarithmic Error. Results demonstrate that t h e LSTM model predicted better forecasting in comparison to all considered time-series forecasting models.

ACKNOWLEDGEMENT

The authors acknowledge the immense help received from the scholars whose articles are cited and included in references to this manuscript. The authors are also grateful to authors/editors/publishers of all those articles, journals and books from where the literature for this article has been reviewed and discussed.

Authors' Contributions

Yash Srivastava, was involved in conceptualization, resources, methodology, software, investigation, writing—original draft, writing—review and editing and visualization. **Shikhar Bhardwaj** supported conceptualization, resources, methodology, software, formal analysis, writing—original draft, writing—review and editing and visualization. **Parvathi. R** contributed to validation, data curation, writing—review and editing, supervision, research guidance.

Conflicts of Interest: We have no conflicts of interest to disclose and the study does not involve human subjects and/or animals dataset

Informed consent: We have used only the open-source dataset which can be accessed by the public. No clinical, Human, or Animal dataset in real-time is not used in this research work.

REFERENCES

1. Gupta R, Pal SK. Trend Analysis and Forecasting of COVID-19 outbreak in India. MedRxiv 2020;6(1):23-26.
2. Bi Q, Wu Y, Mei S, Ye C, Zou X, Zhang Z, et al. Epidemiology and Transmission of COVID-19 in Shenzhen China: Analysis of 391 cases and 1,286 of their close contacts. Med Rxiv 2020 4(1):32-36.
3. Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman JM, Yan P, Chowell G. Short-term forecasts of the COVID-19 epidemic in Guangdong and Zhejiang, China. J Clin Med 2020;9(2):596.
4. Liu ZX, Zhang DG, Luo GZ, Lian M, Liu B. A new method of emotional analysis based on CNN–BiLSTM hybrid neural network. Cluster Comput 2020;23(4):2901-2913.
5. Swapnarekha H, Behera HS, Nayak J, Naik B. Role of intelligent computing in COVID-19 prognosis: A state-of-the-art review. Chaos Solitons Fractals 2020;1(13):1028-1099.
6. Silverstein WK, Stroud L, Cleghorn GE, Leis JA. Wu JT, Leung K, et al. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet 2020;395:689–697
7. Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. Lancet Infect Dis 2020;20(5):553-558.
8. Chung J, Kastner K, Dinh L, Goel K, Courville A, Bengio Y. A recurrent latent variable model for sequential data. arXiv preprint arXiv:1506.02216.2015.
9. Time Series in Python — Exponential Smoothing and ARIMA processes - Benjamin Etienne. Accessed from https://towardsdatascience.com/time-series-in- python-exponential-smoothing-and-Arima-processes-2c67f2a52788 on Feb 9,2019
10. Illustrated Guide to LSTM's and GRU's: A step by step explanation - Michael Phi. Accessed from https://towardsdatascience.com/illustrated-guide-to-lstms- and-GRU-s-a-step-by-step-explanation-44e9eb85bf21 on Sep 24, 2018
11. Elmousalami HH, Hassanien AE. Day level forecasting for Coronavirus Disease (COVID-19) spread: analysis, modelling and recommendations. arXiv preprint arXiv:2003.07778. 2020 Mar 15.
12. Wu T, Ge X, Yu G, Hu E. Open-source analytics tools for studying the COVID-19 coronavirus outbreak. MedRxiv. 2020. doi: https://doi.org/10.1101/2020.02.25.20027433.
13. Zheng Z, Wu K, Yao Z, Zheng X, Zheng J, Chen J. The prediction for the development of COVID-19 in global major epidemic areas through empirical trends in China by utilizing state transition matrix model. BMC Infect Dis 2020;20(1):1-2.
14. Heymann DL, Shindo N. COVID-19: what is next for public health? Lancet 2020;22(2):542 545.
15. Jia L, Li K, Jiang Y, Guo X. Prediction and analysis of Coronavirus Disease 2019. Lancet 2020;5(4):447-450.
16. Peng L, Yang W, Zhang D, Zhuge C, Hong L. Epidemic analysis of COVID-19 in China by dynamical modelling. MedRxiv. 2020. doi: https://doi.org/10.1101/2020.02.16.20023465
17. Wang S, Wang X, Wang S, Wang D. Bi-directional long short-term memory method based on attention mechanism and rolling update for short-term load forecasting. Int J Elect Pow Ener Syst. 2019;109:470-479.
18. How to Develop LSTM Models for Time Series Forecasting - Jason Brownlee. Accessed from https://machinelearningmastery.com/how-to-develop-lstm- models-for-time-series-forecasting/ on Nov 14,2018\