Forecasting Covid-19 Confirmed Cases Using LSTM and ARIMA Models In Algeria

Sahed Abdelkader[1], Mékidiche Mohammed[2], Kahoui Hacen[3]

[1]Autor correspondente, University Centre of Maghnia , Faculty of Economics, Algeria

E-mail: sahed14@yahoo.fr

[2,3]University Centre of Maghnia, Faculty of Economics, Algeria

***Abstract*— The aim of this research paper is to model the number of confirmed daily COVID-19 cases in Algeria using LSTM and ARIMA methods. Data for the work were collected daily from 25/02/2020 to 15/09/2020.**

From verifying the performance of the proposed model, the number of confirmed COVID-19 cases estimated with the actual value obtained in the testing step was compared. The comparison shows that the LSTM models are more precise than the ARIMA model in the estimates of the confirmed case numbers of Covid-19, based on the RMSE forecast accuracy criterion.

Keywords—COVID-19; Forecasting; ARIMA; LSTM; Algeria

I. INTRODUCTION

The COVID-19 pandemic first appeared in December 2019 in Wuhan, China, causing severe havoc around the world (Tandon, 2020, p. 1). While the emergence of this epidemic was due to linking to a live animal seafood market in Wuhan, this epidemic is due to animal origin, where the infection is transmitted from human-to-human and spreads at a high speed (Roosa, 2020, p. 257). The Chinese government has worked to contain this epidemic so that it does not spread outside its borders by implementing many measures on a large scale to contain it, and within a few weeks the infection spread quickly, and as a result, it failed to contain it and the disease spread outside China and reached Countries of the world (Fanelli, 2020, p. 1). On March 11th 2020, the World Health Organization (WHO) declared the 2019 novel coronavirus as global pandemic (Chimmula, 2020, p. 1). For this reason, many scientists and researchers have worked to find ways to help them know the confirmed cases, the number of deaths, and the number of recoveries to control the spread of this epidemic. Statistical and epidemiological methods are among the useful tools in such cases (Ribeiro, 2020, p. 1), and then, Future forecasting of the daily number of confirmed cases, the number of deaths and the number of recoveries is necessary for decision-makers to develop to give them to the health system to provide the necessary capabilities for this (Ribeiro, 2020, p. 3).

Among these statistical methods used to predict time-series for COVID-19, we find the Automatic Regression Moving Average (ARIMA) approach. The reason for its widespread used The reason it is used so widely is that it can obtain useful statistical properties (Gupta, 2020, p. 1).Machine learning technique has also been used to predict the time series, which in turn has been widely used by researchers in this scope (Kırbaş, 2020, p. 1). One of our machine learning methods is the long-term memory networks (LSTM) that are very popular for predicting COVID-19 (Belkacem, 2020, p. 2).

Algeria, like other countries in the world, was not spared from this epidemic, The first confirmed case in Algeria was officially announced on 25 February 2020 in Ouargla region and he was an Italian national. The patient has been sent to a quarantine station. Two other cases were reported on 01 march 2020 in Blida region in the North of Algeria. The epidemic continues to spread in other regions of the country Since then, the number of confirmed cases of COVID-19 has increased day after day (Boudrioua, 2020, p. 2). the Algerian government-mandated several approaches to eradicate the spread of COVID-19 such as trying to control the source of contagion and reducing the number of contacts between individuals by confinement and isolation, All schools/universities and mosques have been closed. To alleviate the epidemic impact.

Through the above, the problematic of this paper include the following:

How effective is the Long-Term Memory Network (LSTM) compared to ARIMA models to daily forecasting the number of confirmed daily COVID-19 cases in Algeria?

This study aims to evaluate and compare the performance of the LSTM model against the ARIMA model in time series analysis and forecasting of COVID-19 cases in Algeria of infected people, the number of deaths, and the number of people recovering from COVID-19 in the short term. As undercurrent conditions, it is imperative for policymakers and healthcare professionals to develop future plans and be appropriately equipped for conditions that may arise due to the rapid spread of COVID-19. Therefore, an essential part of advance planning in this scenario is forecasting the number of cases in the future.

II. LITERATURE REVIEW

There are many studies that have looked at predicting COVID-19, through the use of different statistical methods, and the studies have reached different conclusions ,

This study (Saba, 2020) , aims to model and forecast the spread of the epidemic in Egypt using autoregressive integrated moving average (ARIMA) and nonlinear autoregressive artificial neural networks (NARANN). The results showed that the forecasted cases match well with the officially reported cases, which may help the Egyptian decision-makers to develop short-term future plans to confront this epidemic.

Study for (Shahid, 2020),The aim of this study, the evaluation of proposed prediction models that includes self-regression integrated moving average (ARIMA), support vector regression (SVR), long-term memory (LSTM), and long-term bi-directional memory (Bi-LSTM) are assessed for time series prediction of confirmed cases, deaths and recoveries in ten major countries affected due to COVID-19. The results showed that (Bi-LSTM) is the best compared to other models and this is according to the performance criteria (MAE and RMSE).

Study for (Kırbaş, 2020),The aim of this study is to use an integral automatic regression moving average (ARIMA), a nonlinear self-regression neural network (NARNN), a short long network, and a term memory approach (LSTM). Using six metrics to determine the most accurate model (MSE, PSNR, RMSE, NRMSE, MAPE and SMAPE) to design confirmed COVID-19 cases in Denmark, Belgium, Germany, France, the United Kingdom, Finland, Switzerland and Turkey, the results showed that the LSTM model is the most accurate compared to other models.

Study for (Khan, 2020),This study aims to use Vector Autoregressive time-series models to predict daily new confirmed cases and 19 deaths from COVID-19 for Pakistan and recover cases for ten days. The results showed that the forecasted model a maximum of 5,363 new cases per day with a 95% confidence interval of 3,013 to 3,013. 8,385 cases on July 3, 167 deaths per day with a 95% confidence interval from 112 to 233 and a maximum cure rate of 4,016 cases per day with a 95% confidence interval of 2,182-6405 in the next 10 days.

Study for (Alzahrani, 2020) ,This study aimed to use the Integrated Mobile Self-Regression Rate (ARIMA) model to predict the expected daily number of COVID-19 cases in the Kingdom of Saudi Arabia in the next four weeks. The results showed that the ARIMA model gave the best performance compared to other models.

Study for (Ardabili, 2020) ,This study aimed to analyze and compare machine learning and soft computing models to predict the outbreak of Covid-19 disease, and by examining a wide range of machine learning models, the results showed that there are two models that presented promising results (for example, the multi-layered perception of MLP, and the fuzzy inference system. On the network (ANFIS), this study finally points to the consideration of machine learning as an effective tool for modelling outbreaks.

Study for (Balah, 2020) ,The aim of this study is to forecasting daily confirmed cases in Algeria and this is by using an ARFIMA model is proposed to forecast new COVID-19 cases in Algeria.the results showed that using the expected The forecasted results obtained by the proposed ARFIMA model can be used as a decision support tool to manage medical efforts and facilities against the COVID-19 pandemic crisis.

III. METHODOLOGY

A. ARIMA Model

Autoregressive Integrated Moving Average Models are among the most used statistical techniques for time series analysis is the acronym for $\mathrm{ARIMA}\ (p, d, q)$, which were originally developed for economic applications (Papastefanopoulos, 2020, p. 3). Whereas, the ARIMA model (p, d, q) is a combination of an automatic regression (AR) model that shows that there is a relationship between a value in the present and a value in the past, added by a random value and the moving average model (MA) indicating a relationship between the value in the present and the residues in the past (Badmus, 2011, p. 172). Hence the parameters p and q refer to the order of the AR and MA models, respectively, and d is the level of difference.The general formula for the automatic regression model is shown in Equation Eq(1) where y_t depends only on its lags (Siami-Namini, 2018, p. 1395):

$$y_t = \theta + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} \qquad (1)$$

Where y_t denotes the current values measured at time t; As for the coefficients of θ and α_i; And p is the autoregressive component.

As for a moving average, y_t depends only on the lagging prediction errors, and takes the following formula as shown in Equation Eq(2) :

$$y_t = \varepsilon_t + \emptyset_1 \varepsilon_{t-1} + \emptyset_2 \varepsilon_{t-2} + \cdots + \emptyset_q \varepsilon_{t-q} \qquad (2)$$

Where y_t denotes the current values measured at time t ; ε_t is the forecast error at time t , $\emptyset_i$ are coefficients; And q is the moving average component.

Since the ARMA model is a combination of the terms AR and MA, it takes the following general form In the case of stationary.

$$y_t = \theta_0 + \sum_{i=1}^{p} \theta_i y_{t-i} + \varepsilon_t - \sum_{j=1}^{q} \emptyset_j y_{t-j} \qquad (3)$$

In the case the ARMA process is dynamic and non-stationary time series,a transformation of the series is provided by Box and Jenkins to make it stable and causes the measured values of the ARIMA model. This is achieved by replacing the measured values y_t with the results of a recursive differencing process ∇dy_t where d is the number of times the differencing process has been applied. The first order differencing can be expressed as (Ho, 2002) :

$$\nabla^d y_t = \nabla^{d-1} y_t - \nabla^{d-1} y_{t-1} \qquad (4)$$

B. Long Short-Term Memory Network (LSTM)

LSTM networks are one type of Recurrent Neural Network (RNN). Which works to overcome the problem of not storing previous data in the memory cell, that is, retrieving values from previous stages, for future use (Sagheer, 2019, p. 205). It has been suggested by Hochreiter and Schmidhuber (Hochreiter, 1997, p. 1735)and is an evolution of RNN, in order to address problems of the drawbacks of the RNN by adding additional interactions per module (or cell). It also has, the ability to learning long-term dependencies and remembering information for long periods of time; In addition, the repeating module has a different structure. Instead of a single neural network like standard RNN, it has four layers that interact with a unique communication method (Masum, 2018, p. 6).

As for the calculation of the states of LSTM cells, they are as follows:

The first gate in the LSTM unit is the forget gate f_t, which determines how much information is kept from the last state c_{t-1} . The forget state at time t is formulated as (Le, 2019, p. 8):

$$f_t = \sigma(W_f[h_{t-1}, X_t] + b_f) \qquad (5)$$

Where $\sigma(.)$ denotes the sigmoid activation function,f_t , h_{t-1}, and $b_f$stand for the forget gate vector at time t, the output vector (also the state-h vector) at time $t - 1$, and the bias of the forget gate at time t, respectively; $W_f$is the weight matrix of the forget gate. The following step is deciding and storing information from the new input (X_t) in the cell, state as well as to update the cell state. This step contains two parts, the sigmoid layer and second the tanh layer. First, the sigmoid layer decides whether the new information should be updated or ignored (0 or 1) we mean by this, On the basis of its inputs, if the forgotten gate outputs '1', it indicates "reserve", and if it results "0", it indicates 'discard' data in a cell (Alazab, 2020, p. 174), and second, the tanh function gives weight to the values which passed by, deciding their level of importance (−1 to 1). The two values are multiplied to update the new cell state. This new memory is then added to an old memory C_{t-1} resulting in C_t.

$$i_t = \sigma(W_i[h_{t-1}, X_t] + b_i) \qquad (6)$$

$$N_t = \tanh(W_n[h_{t-1}, X_t] + b_n) \qquad (7)$$

$$C_t = C_{t-1} f_t + N_t i_t \qquad (8)$$

C_{t-1} and C_t denote the cell states at time $t - 1$ and t, While $W_i$and b_i refer to the weight and bias matrix of the input gate , respectively, of the cell state. It can i_t be seen that it has a syntax similar to f_t.Tanh (.) also denotes the tanh activation function; while $W_n$and b_n refer to the weight matrix and the bias of the current gate, respectively.

In the final step of the LSTM unit is to calculate how much information can eventually be treated as the output. Another control gate is chosen as the output gate O_t Where, $w_o$is weight matrices and b_o is the and bias matrice, of the output gate :

$$O_{t=}\sigma * (W_0[h_{t-1}, X_t] + b_0) \qquad (9)$$

Since gates control the information flow by performing an element-wise product, the final output of LSTM h_t is defined by is determined by multiplying by the new values created by the tanh layer from the cell state C_t,with a value ranging between −1 and 1.

$$h_t = O_t \tanh(C_t) \qquad (10)$$

C. Model selection criteria

As it is known, the accuracy of a model can be tested by comparing the actual values with the predicted values, and in this study we apply two performance criteria: the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) are employed, were applied to test the predictive accuracy of the ARIMA model and LSTM model. You write in the following mathematical formulas (Tian, 2018, p. 7):

$$\text{MAE} = \frac{1}{N}\sum_{t=1}^{N}|\hat{y}_t - y_t| \qquad (11)$$

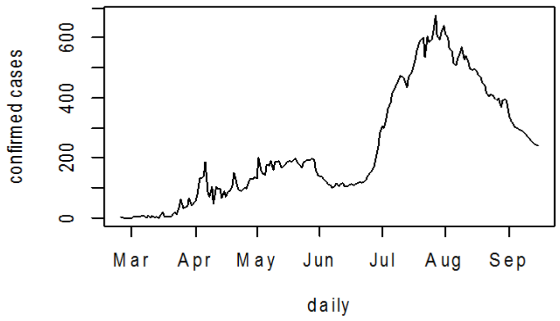
$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(\hat{y}_t - y_t)^2} \qquad (12)$$

IV. RESULTS

A. ARIMA Model

Figure 1 shows the evolution of the Number of daily confirmed COVID-19 cases, during the period 02/25/2020 to 09/15/2020, as the number of confirmed cases continues to rise until reaching the highest value where the number of confirmed cases 675 in 07/27/2020, and then began to decrease, and the average cases is estimated at 234.

Figure01: Number of daily confirmed COVID-19 cases in Algeria: 02/25/2020 to 09/15/2020



The study data are split into two parts, the training set uses observations from 02/25/2020 to 08/31/2020 for the training of the model, and the remaining data were used to test the accuracy of the forecast of the proposed model.

Table 1 shows that the time series under study is not stationary at level, depending on the ADF, PP and KPSS test. For this reason, the difference method was used, which indicates that the time series has become stationary.

Table 01. Result of Unit Root Tests

Test	Level	Decision	Diff	Decision
ADF	0.9163	No stationary	0.01	stationary
PP	0.8814	No stationary	0.01	stationary
KPSS	0.01	No stationary	0.1	stationary

In order to define the model, the simple and partial autocorrelation function was examined so, looking at Figure 2, we find that the most accurate model is ARIMA(0,1,1).

Figure 02: Correlogram at first difference

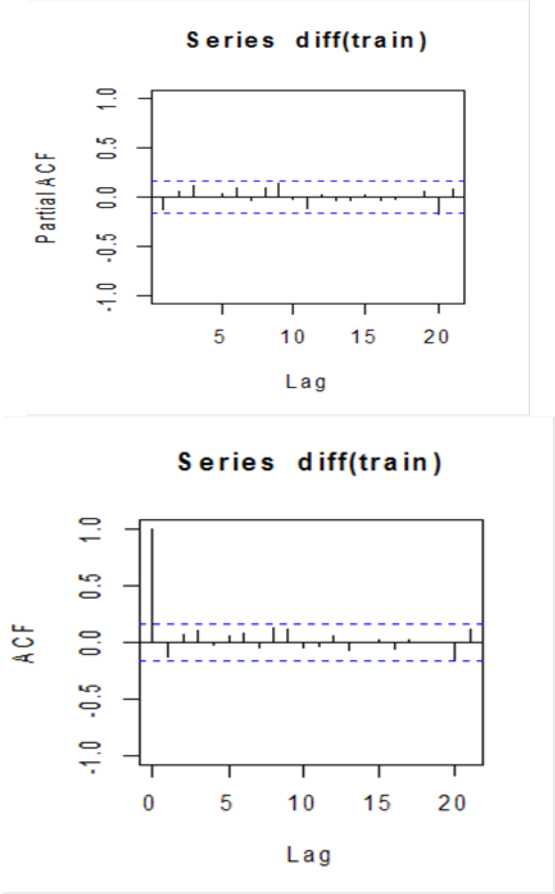
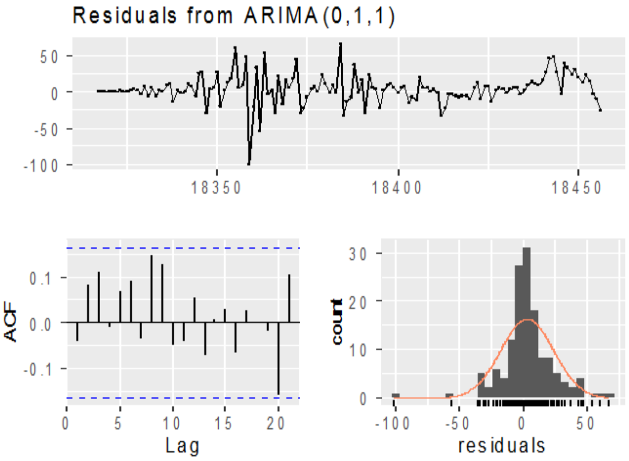


Figure 03: Residuals diagnostics for ARIMA(0,1,1) model



The diagnostic step is taken following the estimate of the ARIMA (0,1,1) model using the maximum likelihood approach.

The Residuals was tested for a white noise operation. It was tested. The ACF of the residuals as shown in Figure 3 indicates that all autocorrelations of the samples fall within the confidence limits of 95 percent for all delays. The histogram also traces a symbolic of normality for a bell shaped distribution.

B. LSTM Model

The tensorflow and keras were used to implement the LSTM models.

The Parameters for LSTM include one input layer, optimization algorithm is Adam, loss function is mean squared error estimated at 9.8401e-04, and maximum number of iterations is 100.

Table 4 summarizes the result of the comparison between the ARIMA and LSTM model in the training and testing part.

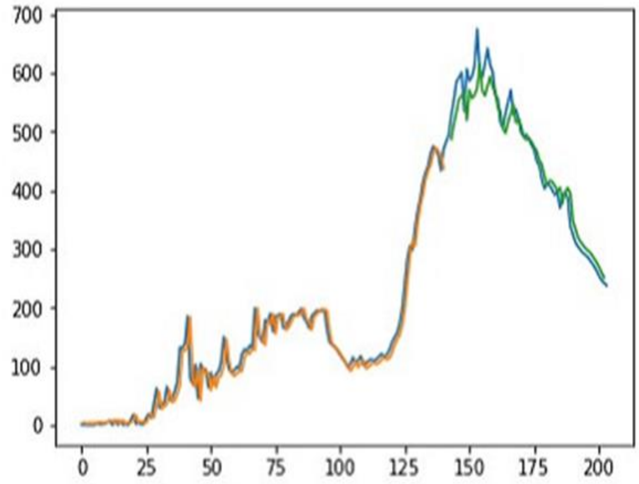
The results are reported in terms of RMSE, as evidenced by the results that LSTM is better than ARIMA in terms of the criterion of prediction accuracy used, and thus the appropriate LSTM model in predicting the number of confirmed cases of Covid-19 virus.

Table 02. RMSE for ARIMA, LSTM on train and test portion of time series

Train/Test	ARIMA	LSTM
	RMSE	RMSE
Train	20.81	21.48
Test	123.22	30.42

As shown in Figure 4, we can say that the LSTM model did an excellent job of simulating the flow of the time series under study in the testing part.

Figure 04: Comparison between the target and output on the LSTM model



V. CONCLUSION

Eventually, in this study, two different models (LSTM & ARIMA) were used to estimate the number of confirmed COVID-19 cases. The LSTM model is based on Tensorflow and Keras, the LSTM model Parameters include a single input layer, the Adam optimization algorithm, a mean loss function and 100 iterations.

The comparison shows that in the estimation of the daily confirmed case numbers of Covid-19, the LSTM models are more accurate than the ARIMA model, and this was based on the RMSE forecast accuracy criterion. However, we recommend reforming the health sector in Algeria to avoid harmful consequences of this pandemic for the economy and society over time.

REFERENCES

1. Tandon, H., Ranjan, P., Chakraborty, T., & Suhag, V. (2020). *Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future*. arXiv preprint arXiv:2004.07859.p1.
2. Roosa, K., Lee, Y., Luo, R., Kirpich, A., Rothenberg, R., Hyman, J. M., ... & Chowell, G. (2020). *Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020*. Infectious Disease Modelling, 5, 256-263.p257.
3. Fanelli, D., & Piazza, F. (2020). *Analysis and forecast of COVID-19 spreading in China, Italy and France*. Chaos, Solitons & Fractals, 134, 109761.p1.
4. Chimmula, V. K. R., & Zhang, L. (2020*). Time series forecasting of COVID-19 transmission in Canada using LSTM networks.* Chaos, Solitons & Fractals, 109864.p1.
5. Ribeiro, M. H. D. M., da Silva, R. G., Mariani, V. C., & dos Santos Coelho, L. (2020). *Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. Chaos*, Solitons & Fractals, 109853.p1-3.
6. Gupta, R., & Pal, S. K. (2020). *Trend Analysis and Forecasting of COVID-19 outbreak in India*. medRxiv.p1.
7. Kırbaş, İ., Sözen, A., Tuncer, A. D., & Kazancıoğlu, F. Ş. (2020). *Comperative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches.* Chaos, Solitons & Fractals, 110015.p1.
8. Belkacem, S. (2020). *COVID-19 data analysis and forecasting: Algeria and the world*. arXiv preprint arXiv:2007.09755.p2.
9. Boudrioua, M. S., & Boudrioua, A. (2020). *Predicting the COVID-19 epidemic in Algeria using the SIR model*. medRxiv.p2.
10. Saba, A. I., & Elsheikh, A. H. (2020*). Forecasting the prevalence of COVID-19 outbreak in Egypt using nonlinear autoregressive artificial neural networks*. Process Safety and Environmental Protection.
11. Shahid, F., Zameer, A., & Muneeb, M. (2020). *Predictions for COVID-19 with Deep Learning Models of LSTM, GRU and Bi-LSTM. Chaos*, Solitons & Fractals, 110212.
12. Khan, F., Saeed, A., & Ali, S. (2020). *Modelling and Forecasting of New Cases, Deaths and Recover Cases of COVID-19 by using Vector Autoregressive Model in Pakistan*. Chaos, Solitons & Fractals, 110189.
13. Alzahrani, S. I., Aljamaan, I. A., & Al-Fakih, E. A. (2020). *Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions.* Journal of infection and public health, 13(7), 914-919.
14. Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R., Reuter, U., ... & Atkinson, P. M. (2020). *Covid-19 outbreak prediction with machine learning*. Available at SSRN 3580188.
15. Balah, B., & Djeddou, M. (2020). *Forecasting COVID-19 new cases in Algeria using Autoregressive fractionally integrated moving average Models (ARFIMA*). medRxiv.
16. Papastefanopoulos, V., Linardatos, P., & Kotsiantis, S. (2020). COVID-19: *A Comparison of Time Series Methods to Forecast Percentage of Active Cases per Population*. Applied Sciences, 10(11), 3880.p3.

17. Badmus, M. A., & Ariyo, O. S. (2011*). Forecasting cultivated areas and production of maize in Nigerian using ARIMA Model*. Asian Journal of Agricultural Sciences, 3(3), 171-176.p172.

18. Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2018, December). *A comparison of ARIMA and LSTM in forecasting time series*. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1394-1401). IEEE.p1395.

19. Ho, S. L., Xie, M., & Goh, T. N. (2002). *A comparative study of neural network and Box-Jenkins ARIMA modeling in time series prediction*. Computers & Industrial Engineering, 42(2-4), 371-375.

20. Sagheer, A., & Kotb, M. (2019). *Time series forecasting of petroleum production using deep LSTM recurrent networks.* Neurocomputing, 323, 203-213.p205.

21. Masum, S., Liu, Y., & Chiverton, J. (2018, June). *Multi-step time series forecasting of electric load using machine learning models*. In International Conference on Artificial Intelligence and Soft Computing (pp. 148-159). Springer, Cham.p6.

22. Alazab, M., Awajan, A., Mesleh, A., Abraham, A., Jatana, V., & Alhyari, S. (2020). *COVID-19 Prediction and Detection Using Deep Learning.* International Journal of Computer Information Systems and Industrial Management Applications, 12, 168-181.p174.

23. Tian, C., Ma, J., Zhang, C., & Zhan, P. (2018). *A deep neural network model for short-term load forecast based on long short-term memory network and convolutional neural network*. Energies, 11(12), 3493.