SARIMA-LSTM COMBINATION FOR COVID-19 CASE MODELING

IMAM TAHYUDIN[1], RIZKI WAHYUDI[2*] AND HIDETAKA NAMBO[3]

[1]*Department of Information System, Universitas Amikom Purwokerto, Purwokerto, Indonesia*
[2]*Department of Informatics, Universitas Amikom Purwokerto, Purwokerto, Indonesia*
[3]*Artificial Intelligence Laboratory, Kanazawa University, Japan*

[*]*Corresponding author: rizkiw@amikompurwokerto.ac.id*

(Received: 4th July 2021; Accepted: 3rd December 2021; Published on-line: 4th July 2022)

ABSTRACT: The study of SARIMA method in combination with LSTM is interesting to do. This combination method can be convincing and significant because the data collected is numerical and saved based on time. In addition, the proposed method can anticipate datasets, either linear or non-linear. Based on several previous studies, the SARIMA method has the advantage of completing linear datasets while the LSTM method excels in achieving non-linear datasets. Also, both methods have been shown to have an accuracy value compared to some other methods. This study tried to combine the two through several stages of the first stage of applying the SARIMA method using fit datasets (linear data) then residual Dataset (non-linear data) analysed using the LSTM method. The result of the combination methods will be checked for the accuracy value. This research will be compared by using SARIMA and LSTM methods separately. The Dataset used as a trial is COVID-19 patient data in the United States. The results showed that the combination of SARIMA-LSTM method is better than either SARIMA or LSTM alone with RMSE of 0.33905765 and MAE of 0.29077017.

ABSTRAK: Gabungan kaedah kajian SARIMA dengan LSTM adalah menarik untuk dikaji. Gabungan kaedah ini meyakinkan dan penting kerana data yang dikumpulkan bersifat numerik dan disimpan berdasarkan waktu. Selain itu, kaedah yang diusulkan ini dapat menerima set data, samada berkadar langsung atau tidak langsung. Berdasarkan beberapa penelitian sebelumnya, kaedah SARIMA mempunyai faedah dalam melengkapi set data linear, sedangkan kaedah LSTM berguna dalam mencapai set data tidak-linear. Tambahan, kedua-dua kaedah ini terbukti memiliki nilai ketepatan lebih baik berbanding beberapa kaedah lain. Kajian ini cuba menggabungkan keduanya melalui beberapa tahap. Tahap pertama mengunakan kaedah SARIMA secara set data (data linear) kemudian baki set data (data tidak-linear) dianalisa menggunakan kaedah LSTM. Dapatan dari gabungan kedua-dua kaedah tersebut akan diperiksa nilai ketepatannya. Kajian ini akan dibandingkan melalui kaedah SARIMA dan LSTM secara berasingan. Set data yang digunakan adalah merupakan data pesakit COVID-19 dari Amerika Syarikat. Dapatan kajian menunjukkan gabungan kaedah SARIMA-LSTM memiliki nilai ketepatan yang lebih baik berbanding kaedah SARIMA secara berasingan, dan LSTM dengan RMSE adalah sebanyak 0.33905765 dan MAE sebanyak 0.29077017.

KEYWORDS: *SARIMA; LSTM; SARIMA-LSTM; COVID-19 patients*

1. INTRODUCTION

The term SARIMA designates a seasonal autoregressive integrated moving average. One of the time series method's topics is this model. To handle problems involving time-run results, time series are frequently used. The time-series methodology is used in a

variety of domains, including the economics and financial turnaround results at hospitals [1-3].

A discussion of the time series method was held to complete the modelling of potential bioelectric plant data. Using autoregressive (AR), moving average (MA), and ARIMA models are one example. However, some of these models have not produced the best value. The average error rate for mean square error (MSE) and the mean absolute error (MAE) continues to be strong. Although the average prediction accuracy is still about 75% [4-6]. Therefore, this study aims to increase accuracy by using another time series model, SARIMA. In this study, the SARIMA method was combined with the LSTM method.

The combination of SARIMA with other methods is proven to have better accuracy results. Among them is the combination of the SARIMA method with other methods, including the SVM method, to predict the production value of the machine industry in Taiwan [7]. The results showed that SARIMA hybrid accuracy with SVM is better than with each method. Other studies used hybrid ARIMA with ANN for forecasting pollution index in cities throughout Southeast Asia and further research used the same approach to predict tourists coming at Minangkabau international airport [8]. The results showed the accuracy value of hybrid methods is better. Subsequent research compared SARIMA and ANN methods to predict power absorption in Turkey's electricity users [9]. After 12 weeks, the results showed that the ANN method's MAPE value was 1.8% better than SARIMA because it had a MAPE of 2.6%. However, in certain conditions, such as the time after a holiday, the result is the opposite. Another study combined SARIMA with SVM, and then in analysis using clustering [10], this research was used to predict passengers at northern Iranian stations. The result is a better mix of these approaches than the individual methods. As a result, the integrated approach outperforms the respective processes. Additional research on the combination of SVR-SARIMA models was done for tourist forecasting [11], for the best model's determination using the decision support system PROMETHEE II. The result is the same combination method is better.

Thus, this study combined the SARIMA method with LSTM. The SARIMA model successfully predicts a person's position for linear data set type better than the deep learning method [12], and also the SARIMA model has been tested with high accuracy of about 80% [13]. According to research, the Long-Short Term Memory (LSTM) Recurrent Neural Network on Workload Forecasting Models for Cloud Datacentres has generated empirical results. The proposed method achieves high accuracy in prediction by reducing average squared errors by up to 3.17 x 10^{-3} [14]. Therefore, we use both methods because both can solve problems for linear and non-linear data sets. In addition, this study will compare the SARIMA-LSTM combination with each method separately.

2. METHOD

2.1 ARIMA Model

ARIMA is a term derived from its parts: autoregressive (AR), integration (I), and moving average (MA) shape. In general, the ARIMA models are classified into two types, namely non-traditional (non-seasonal) ARIMA and Seasonal ARIMA models [15-17]. The ARIMA model is as follows: ARIMA (p,d,q). p represents the sum of AR values, d is the value of integration (I), and q is the MA value. In general, the ARIMA model (p,d,q) can be seen from the model as follows:

$$(1 - \phi_1 B.... \phi_p B^p)(1 - B)^d Y_t = c + (1 + \theta_1 B..... + \theta_q B^q)\, et \quad (1)$$

There are 3 main components in the model, the first being AR (p),

$$(1-\phi_1 B.... - \phi_p B^p) \quad (2)$$

The second is differentiation through I (d),

$$(1 - B)^d Y_t \quad (3)$$

The third was indeed MA (q),

$$(1 + \theta_1 B..... + \theta_q B^q)e_t \quad (4)$$

c, on the other side, is a constant value.

2.2 Seasonal ARIMA Model

The seasonal ARIMA, or SARIMA model, is a model or shape that repeats itself at regular intervals. For stationary datasets, seasonality can be detected from the ACF plot. If the ACF visualization shows seasonal patterns, it will be done with a different solution [18-20]. In general, the seasonal ARIMA equation is shown in eqn. 5.

$$ARIMA(p,d,q)(P,D,Q)^s \quad (5)$$

where (p,d,q) is the non-seasonal ARIMA model index, while (P, D, Q) is the seasonal ARIMA model, and S is the number of periods on the seasonal model.

For example, if ARIMA (1,0,0), then the model follows the following eqns. 6 and 7:

$$(1 - \phi_1 B)Y_t = c \quad (6)$$

where is $BY_t = Y_{t-1}$. So

$$Y_t = c + \phi_1 Y_{t-1} \quad (7)$$

To detect seasonal datasets, there are several chart techniques including sequential plots, seasonal plot subseries, multiple box plots, and autocorrelated plots. The study will use autocorrelated schemes to detect seasonality. One of the solutions for this autocorrelation plot is to use seasonal differential operators.

2.3 LSTM

Long-short-term Memory (LSTM) is a form of RNN that consists of a collection of cells with features that allow them to memorize data sequences. Data streams are captured and stored in cells. The cell then connects one module from the past to another, allowing data to be transmitted from several previous instances to the present. The data in every cell can be rejected, screened or started adding as a result of the gates in every cell in preparation for the cells that come after [19,21].

The gates focus on a neural network with sigmoidal shape layers, and the active cells either transfer data or discard it. Each sigmoid layer generates a number between 0 and 1, indicating the sum of each data segment that must be permitted in every cell. More precisely, The estimated low value assumes that "nothing should be allowed to pass"; while forecast one shows that "let it all pass". Every LSTM has three types of gates that regulate a state for every cell:

1) **Forget Gate** produces a value between 0 and 1, with 1 denoting success. "fully save this"; whereas 0 says "ignore this."
2) **Memory Gate** The sigmoid layer, following either by the tanh layer, determines which of the cell's most recent data must be kept. The first sigmoid layer, known as

the "doorway layer," selects which values to modify. The tanh layer then generates a new candidate value vector, which can be added to the state.

3) **The Output Gate** determines what to make from each cell. The final value will be determined by the cell state as well as newly added filtered data.

If the distance is vast, the RNN will be unable to predict the next result. Consider the following text: "I go to work every day" and "I work hard at the office." The location's name is the next possible word for current knowledge, but deciding what kind of location to use is difficult. Since there was some related knowledge in the previous period, RNN cannot learn to relate information. As a result, LSTM is a solution for overcoming these flaws.

LSTMs can study long-term dependencies. Remembering information for a long time is the default behavior. Some of the equations show this module as follows [22], [23].

$$r_t = sigm(W_{xr}x_t + W_{hr}h_{t-1} + b_r) \quad (8)$$

$$z_t = sigm(W_{xz}x_t + W_{hz}h_{t-1} + b_z) \quad (9)$$

$$\tilde{h}_t = tanh(W_{xh}x_t + W_{hh}(r_t \odot h_{t-1}) + b_h) \quad (10)$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t \quad (11)$$

2.4 Combination of SARIMA Method with LSTM

The steps of this combination are described in Fig. 1.

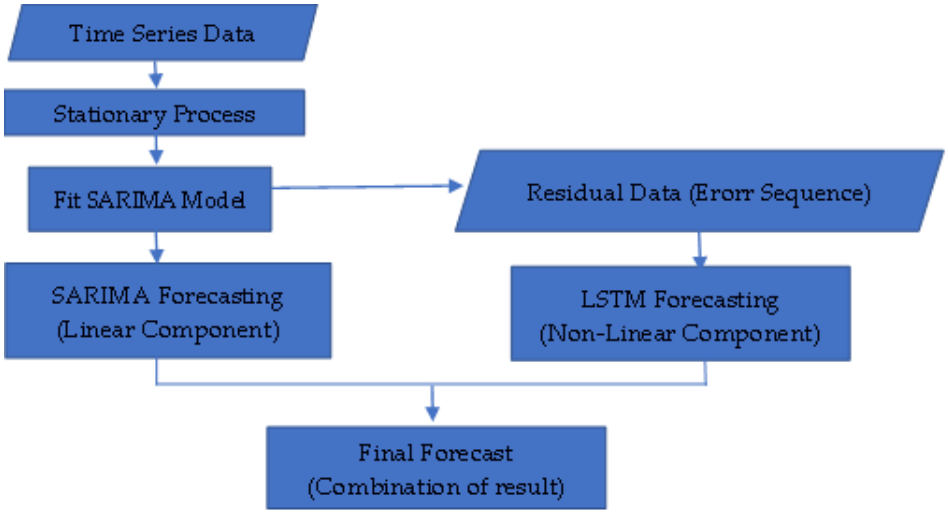

Fig. 1: Combination SARIMA-LSTM Model of COVID-19 patient.

Based on Fig. 1, the combination of the two methods is as follows: (1) Prepare datasets; in this case, the Dataset used is the COVID-19 patient death data in a country affected by COVID-19. (2) Analysis process using the SARIMA model. The method includes the identification process. This process is checked using the Box-Jenkins process. That is the process of determining static data. The seasonal Dataset is then examined by looking at the ACF and PACF values. (3) If the Dataset is seasonal, then the next step determines the best SARIMA model. (4) SARIMA Fit model is forwarded for the forecasting process for linear components. While residual data is used for the forecasting process using the LSTM method for non-linear components. (5) The last

step is the combined result of forecasting the two methods, further measuring MSE values.

3. RESULTS AND DISCUSSION

3.1 SARIMA Model Implementation

In implementing the SARIMA method, COVID-19 patients who died based on gender in the United States were used. The original data used as many as 1008 datasets but because many values were empty (missing value), the empty data was deleted. The net data was 711 datasets. The analysis process used R and Minitab software.

3.1.1 Identification

As a dataset trial, male COVID-19 patients' datasets were used. A plot of male COVID-19 patients who died in the United States is shown in Fig. 2.

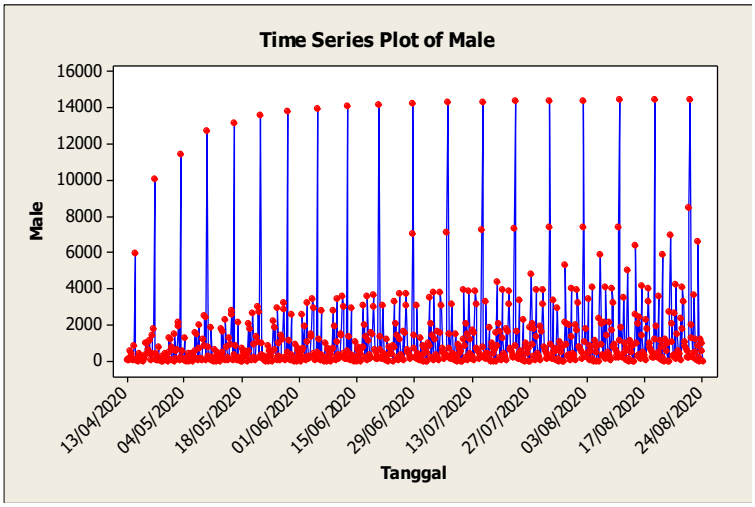

Fig. 2: Visual Dataset of male COVID-19 patients who died in the USA.

Following stationary check dataset of male COVID-19 patients.

a) Stationary check of variety with Cox Box test Results obtained as follows:

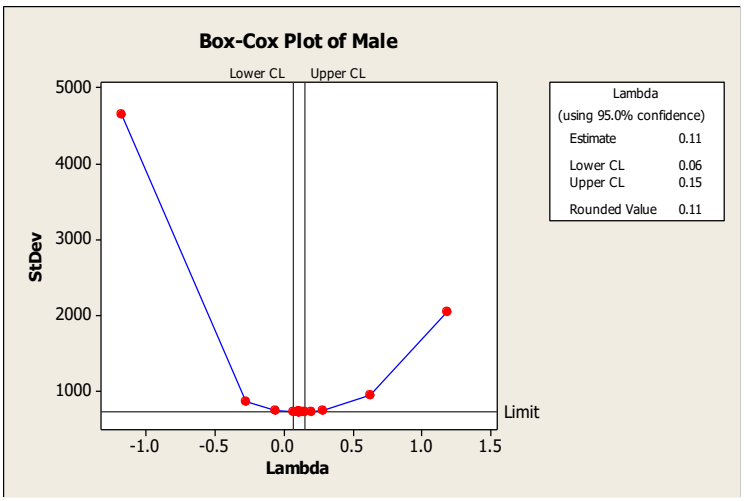

Fig. 3: Visual Dataset of male COVID-19 patients who died in the USA.

Because the value is still 0.11 should be worth 1. So, the transformation process is carried out as follows:

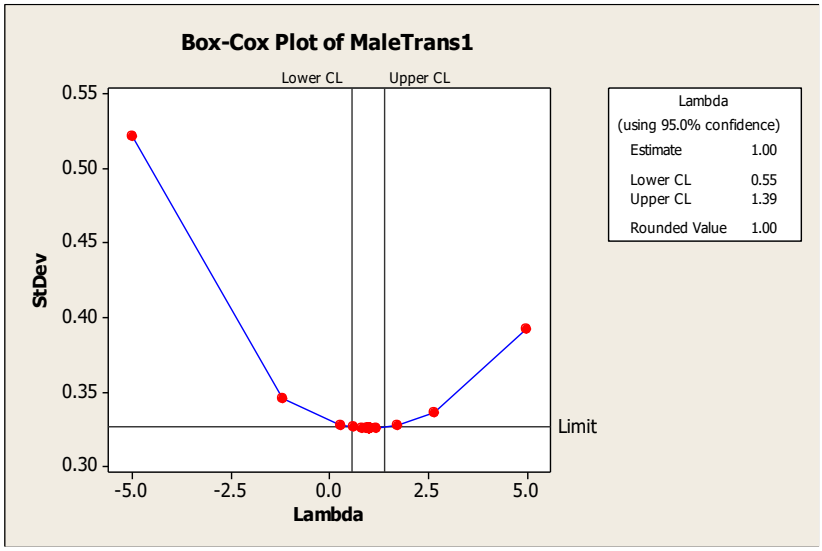

Fig. 4: Cox Box Transformation.

Since the lambda value = 1 is stationary. It then checked stationary against the average by looking at its ACF and PACF scores.

b) Stationary checks against averages by checking ACF and PACF values

Based on ACF and PACF lag, 1-3 images are still inside the significance interval. Then it has been declared stationary against the average.

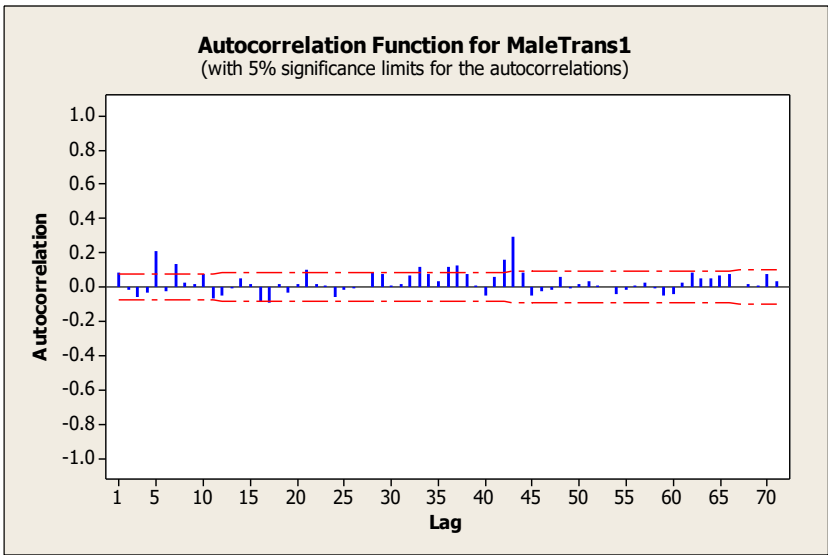

Fig. 5: ACF.

https://doi.org/10.31436/iiumej.v23i2.2134

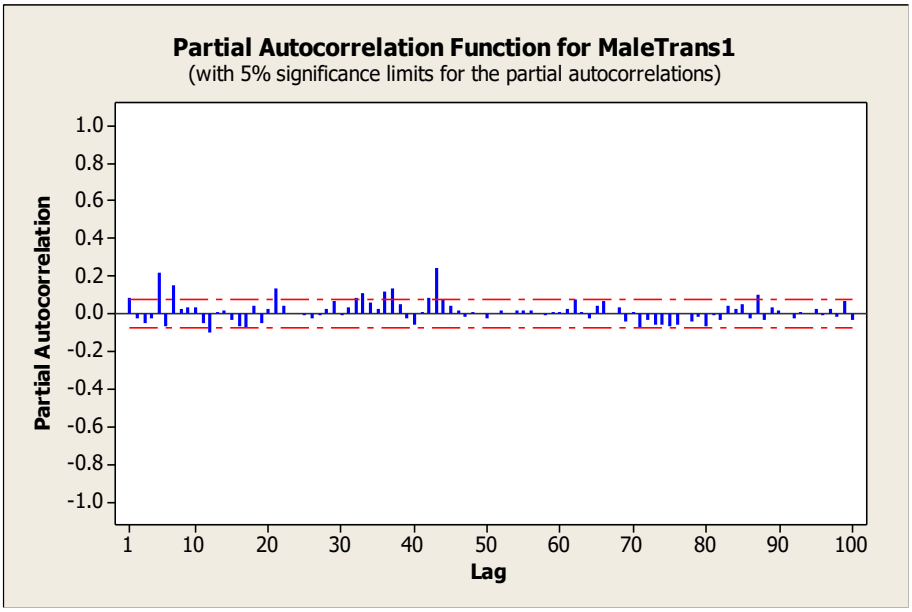

Fig. 6: PACF.

Based on Fig. 12, it appears that there is a trend. That is, there is an increase in the number of COVID-19 patients who die per 10 datasets. Therefore, the identification process is carried out further by differencing

3.1.2 Estimation

In this process, an analysis is carried out based on the sex of the male.

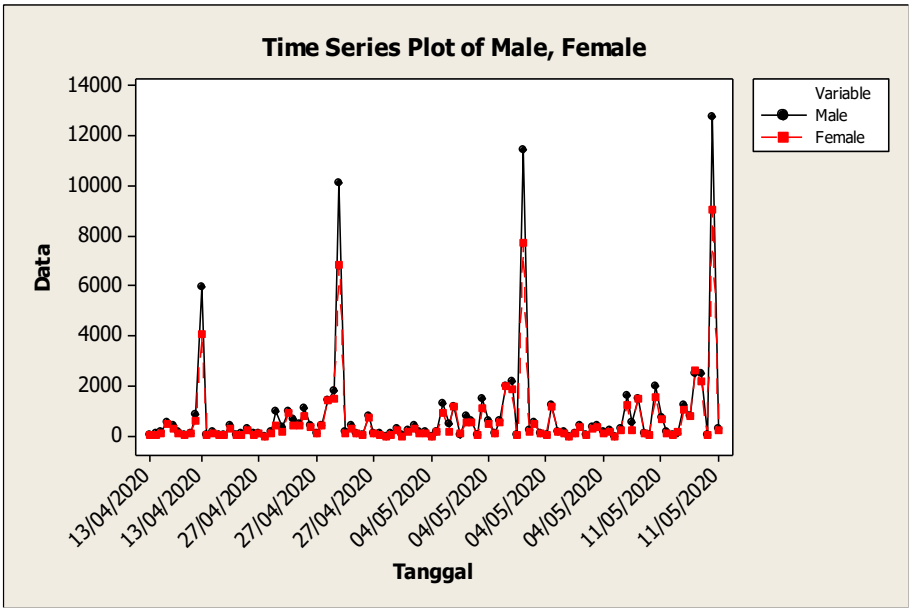

Fig. 7: DATASET of COVID-19 patients in the USA of the male gender.

Once the first differencing process trend is known lag=1, The model is formed with a non-seasonal model first by inspecting the graph autocorrelation function (ACF) and partial autocorrelation function (PACF). The following ACF and PACF charts were obtained:

https://doi.org/10.31436/iiumej.v23i2.2134

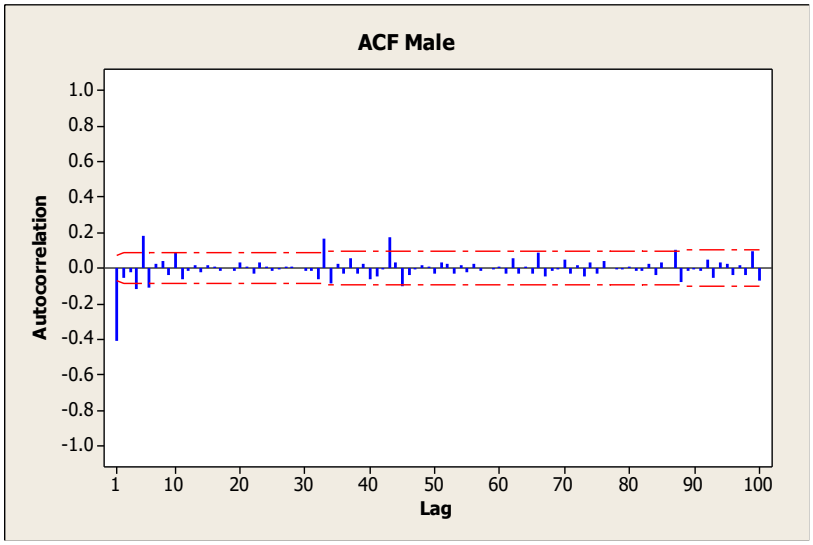

Fig. 8: ACF graph of male COVID-19 patients in America non-seasonal ARIMA models

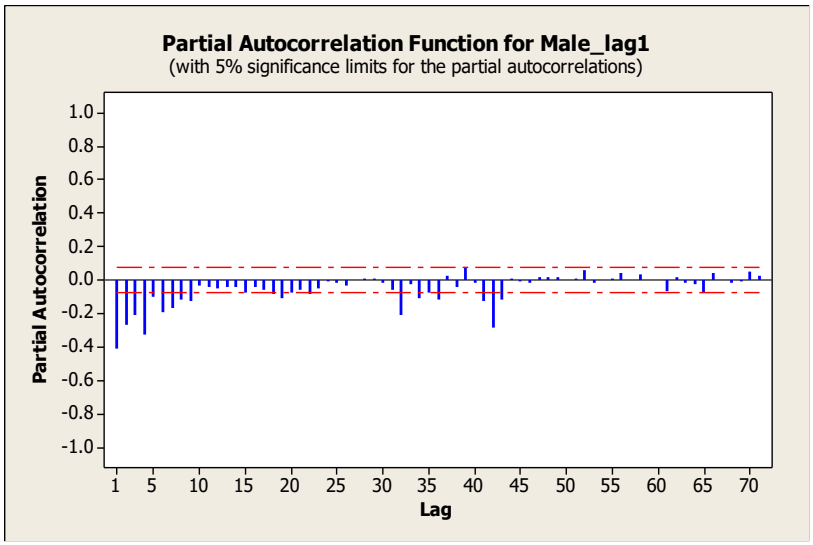

Fig. 9: PACF graph of male COVID-19 patients in America non-seasonal ARIMA model.

The ACF chart shows the dying down as much as five lags. Meanwhile, the PACF chart shows the cut-off pattern. Thus, the non-seasonal ARIMA model formed is ARIMA (5,1,0). They are furthermore checking the non-seasonal ARIMA model.

The step to determine the seasonal ARIMA model is the same as that performed to find the best model in non-seasonal ARIMA by determining its ACF and PACF charts. The seasonal value determines the difference. In this case, it is 10 because for every 10 datasets, there is a significant increase in COVID-19 patient deaths.

3.1.3 Model Evaluation

At this stage, checking the error value and other values. Based on the output of the auto_arima, the evaluation value as follows.

SARIMA $(2,1,2)(0,0,2)^{12}$

Next, check the error value and accuracy. Here are the evaluation values obtained:

https://doi.org/10.31436/iiumej.v23i2.2134

AIC = 340.08

BIC = 374.8

RMSE = 0.44236361

MAE = 0.33180744

These values are the best value when compared to other models.

3.1.4 Forecasting

Based on Fig. 10, the number of COVID-19 patients who died of male gender in the USA, in general, decreased for the following forecasting result. The average difference is about 500 people.

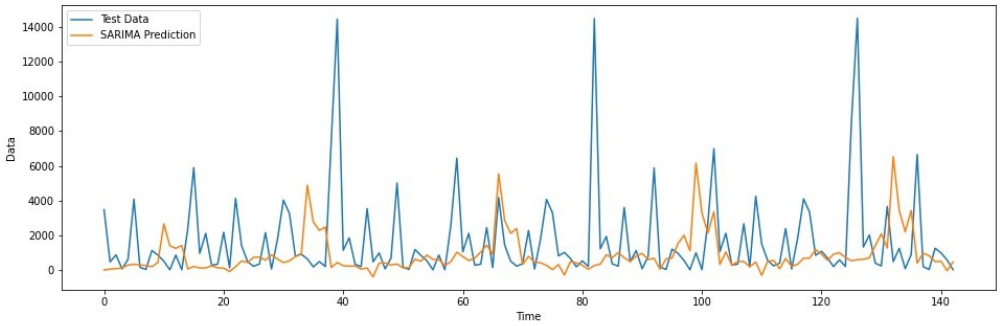

Fig. 10: Male COVID-19 patient forecasting graph using SARIMA.

3.2 Implementation of LSTM and Combination of SARIMA-LSTM method

This analysis used the parameters of batch size of 100, the look_back of seven data to previous, and the epoch for learning of 100. Furthermore, for the forecasting process we split the data set with the composition of 80% training data and 20% testing data from the same data: male COVID-19 patients in the USA. Based on the analysis using LSTM obtained an RMSE value of 0.35847506 and MAE of 0.29837463. Based on RMSE and MAE results, the values are smaller than the SARIMA result, so it can be said that LSTM is better than SARIMA. From Fig 11, it appears that the number of COVID-19 patients who died on average tends to decrease.

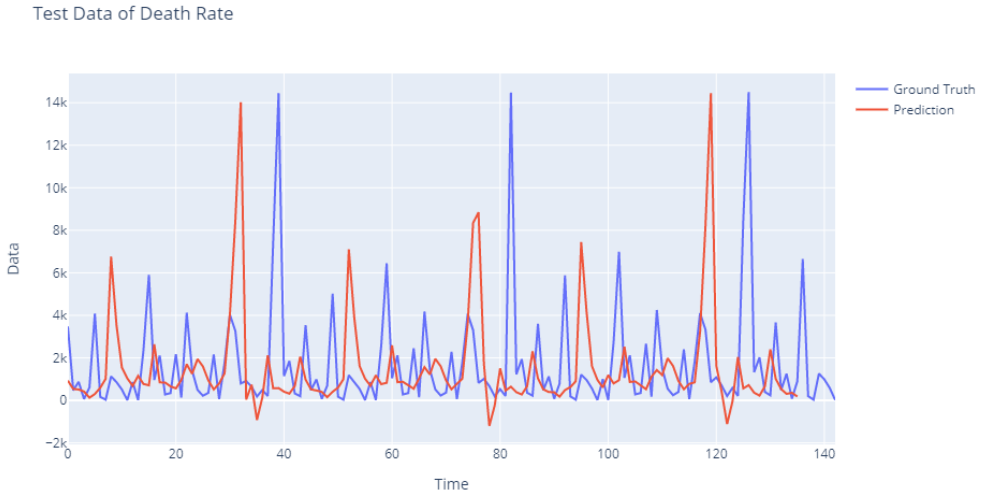

Fig. 11: Graph forecasting male COVID-19 patients using LSTM method.

For the next test, the SARIMA method combined with LSTM was done to determine how much accuracy through RMSE and MAE value obtained and prediction result. Based on the calculation resulted that an RMSE value was 0.33905765 and MAE was 0.29077017.

Those results presented that the RMSE and MAE of combination SARIMA-LSTM was better than the SARIMA and LSTM methods (Table 1). The results of combination SARIMA and LSTM are seen in Fig 12. This figure performed that the result is the best result because the predicted data is almost similar with real data. The number of COVID-19 suspected deaths decreased to near-zero. In addition, the table of comparison of RMSE and MAE are presented in Table 1. This result presented that the combination of SARIMA and LSTM is the best method for predicting the death of COVID-19 patients in the USA.

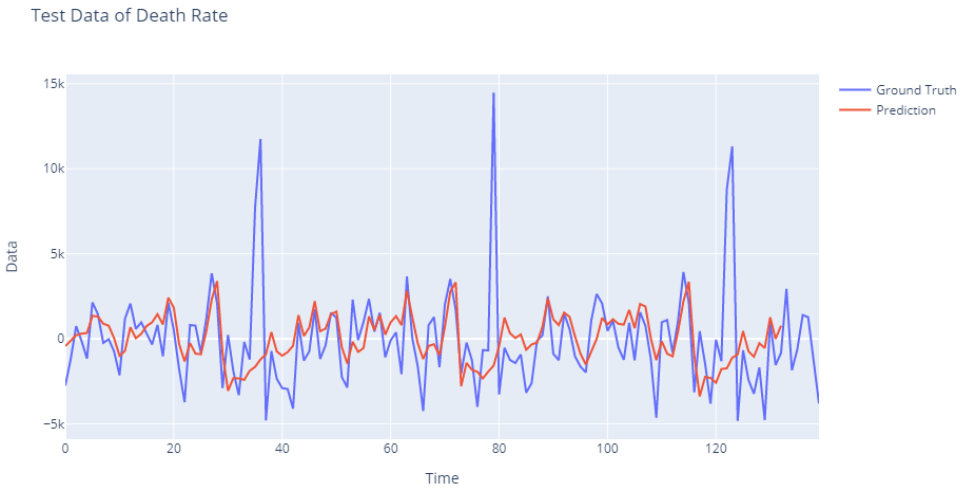

Fig. 12: Male COVID-19 patient forecasting using SARIMA-LSTM combination method

Table 1: The comparison of RMSE and MAE from SARIMA, LSTM, and combination SARIMA-LSTM

Parameters	SARIMA	LSTM	SARIMA-LSTM
RMSE	0.44236361	0.35847506	0.33905765
MAE	0.33180744	0.29837463	0.29077017

4. CONCLUSION

Based on the results of the study, the combination SARIMA-LSTM method is the best one. It performed better than the SARIMA or LSTM methods separately. Based on the results of general predictions using all three methods, there was a decrease in the number of male COVID-19 patients who died in the USA on average. This research has the limitation of not explaining the mortality number of patients in every state within the USA. For future work, the analysis will be carried out using a combination of SARIMA – PARCD methods.

REFERENCES

[1] Davis RA. (2014) Introduction to statistical analysis of time series. Department of Statistics Columbia University, pp. 1-24.

[2] Borkowf CB. (2002) Time-Series Forecasting. Technometrics, 44(2): 194-195. https://doi.org/10.1198/tech.2002.s718.
[3] Schlüter T. (2012) Knowledge discovery from time series (Doctoral dissertation, Universitäts-und Landesbibliothek der Heinrich-Heine-Universität Düsseldorf).
[4] Chen KY, Wang CH. (2007) A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan. Expert Systems with Applications, 32(1): 254-264. https://doi.org/10.1016/j.eswa.2005.11.027
[5] Chi YN. (2021) Time Series Forecasting of Global Price of Soybeans using a Hybrid SARIMA and NARNN Model: Time Series Forecasting of Global Price of Soybeans. Data Science: Journal of Computing and Applied Informatics, 5(2): 85-101. https://doi.org/10.4108/eai.2-8-2019.2290473
[6] Ozozen A, Kayakutlu G, Ketterer M, Kayalica O. (2016) A combined seasonal ARIMA and ANN model for improved results in electricity spot price forecasting: Case study in Turkey. In 2016 Portland International Conference on Management of Engineering and Technology (PICMET) (pp. 2681-2690). IEEE. https://doi.org/10.1109/PICMET.2016.7806831.
[7] Parviz L. (2020) Comparative evaluation of hybrid SARIMA and machine learning techniques based on time varying and decomposition of precipitation time series. Journal of Agricultural Science and Technology, 22(2): 563-578. Retrieved from: http://jast.modares.ac.ir/article-23-26018-en.html
[8] Abellana DPM, Rivero DMC, Aparente ME, Rivero, A. (2020) Hybrid SVR-SARIMA model for tourism forecasting using PROMETHEE II as a selection methodology: a Philippine scenario. Journal of Tourism Futures. https://doi.org/10.1108/JTF-07-2019-0070
[9] Tahyudin I, Nambo H. (2018) Comparison Study of Deep Learning and Time Series for Bioelectric Potential Analysis. In 2018 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE) (pp. 79-83). IEEE. https://doi.org/10.1109/ICITISEE.2018.8720998
[10] Tahyudin I, Nambo H. (2018) SARIMA Model of Bioelectic Potential Dataset. In International Conference on Big Data, Cloud and Applications (pp. 367-378). Springer, Cham. https://doi.org/10.1007/978-3-319-96292-4_29
[11] Kumar J, Goomer R, Singh AK. (2018) Long short term memory recurrent neural network (LSTM-RNN) based workload forecasting model for cloud datacenters. Procedia Computer Science, 125: 676-682. https://doi.org/10.1016/j.procs.2017.12.087
[12] Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M. (2020) Application of the ARIMA model on the COVID-2019 epidemic dataset. Data in Brief, 29: 105340. https://doi.org/10.1016/j.dib.2020.105340
[13] Ceylan Z. (2020) Estimation of COVID-19 prevalence in Italy, Spain, and France. Science of The Total Environment, 729:138817.
[14] Zeroual A, Harrou F, Dairi A, Sun Y. (2020) Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study. Chaos, Solitons & Fractals, 140: 110121
[15] NIST/SEMATECH: Seasonality (2012). http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc443.htm. Accessed 23 September 2020
[16] Qi C, Zhang D, Zhu Y, Liu L, Li C, Wang Z, Li X. (2020) SARFIMA model prediction for infectious diseases: application to hemorrhagic fever with renal syndrome and comparing with SARIMA. BMC medical research methodology, 20(1): 1-7. https://doi.org/10.1186/s12874-020-01130-8
[17] Hamilton JD. (2020) Time series analysis. Princeton university press.
[18] Sherstinsky A. (2020) Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404: 132306. https://doi.org/10.1016/j.physd.2019.132306
[19] Reddy BK, Delen D. (2018) Predicting hospital readmission for lupus patients: An RNN-LSTM-based deep-learning methodology. Computers in biology and medicine, 101: 199-209. https://doi.org/10.1016/j.compbiomed.2018.08.029
[20] Qi J, Liu X, Tejedor J. (2020) Variational inference-based Dropout in recurrent neural networks for slot filling in spoken language understanding. arXiv Preprint arXiv:2009.01003

[21] Li C, Zhao L, Cai B. (2020) Size prediction of railway switch gap based on RegARIMA model and LSTM network. IEEE Access, 8, 198188-198200. https://doi.org/10.1109/ACCESS.2020.3034687

[22] Z. Liu *et al.*, "Entity recognition from clinical texts via recurrent neural network," *BMC Med. Inform. Decis. Mak.*, vol. 17, no. Suppl 2, 2017, doi: 10.1186/s12911-017-0468-7

[23] M. A. Jishan, K. R. Mahmud, A. K. Al Azad, M. S. Alam, and A. M. Khan, "Hybrid deep neural network for bangla automated image descriptor," *Int. J. Adv. Intell. Informatics*, vol. 6, no. 2, pp. 109–122, 2020. https://doi.org/10.26555/ijain.v6i2.499