Prediction of Covid-19 Daily Case in Indonesia Using Long Short Term Memory Method

Faisal Dharma Adhinata[1*], Diovianto Putra Rakhmadani[2]

[1,2]Department of Software Engineering, Faculty of Informatics, Institut Teknologi Telkom Purwokerto, Indonesia
Email: [1*]faisal@ittelkom-pwt.ac.id, [2]diovianto@ittelkom-pwt.ac.id

(Article received: 12 Jan 2021, revised: 5 Feb 2021, accepted: 19 Feb 2021)

Abstract

The impact of this pandemic affects various sectors in Indonesia, especially in the economic sector, due to the large-scale social restrictions policy to suppress this case's growth. The details of the growth of Covid-19 in Indonesia are still fluctuating and cannot be fully understood. Recently it has been developed by researchers related to the prediction of Covid-19 cases in various countries. One of them is using a machine learning technique approach to predict cases of daily increase Covid-19. However, the use of machine learning techniques results in the MSE error value in the thousands. This high number indicates that the prediction data using the model is still a high error rate compared to the actual data. In this study, we propose a deep learning approach using the Long Short Term Memory (LSTM) method to build a prediction model for the daily increase cases of Covid-19. This study's LSTM model architecture uses the LSTM layer, Dropout layer, Dense, and Linear Activation Function. Based on various hyperparameter experiments, using the number of neurons 10, batch size 32, and epochs 50, the MSE values were 0.0308, RMSE 0.1758, and MAE 0.13. These results prove that the deep learning approach produces a smaller error value than machine learning techniques, even closer to zero.

Keywords: Covid-19, machine learning, deep learning, LSTM, MSE.

I. INTRODUCTION

The World Health Organization (WHO) announced that the coronavirus became a global disease outbreak in all countries or what is known as a pandemic [1]. This virus attacks the human respiratory system that can cause pneumonia, and other symptoms such as fever, dry cough, decreased appetite, diarrhea, vomiting, and abdominal pain [2]. This pandemic caused a high mortality rate. Even death cases in Indonesia reached 22,555, with total cases reaching 758,473 as of January 2, 2021 [3]. Recently, cases of people infected with the coronavirus in Indonesia are increasing day by day. Even the number of victims who died reached more than 200 per day. In general, the mechanism for spreading corona cases in Indonesia is not fully known.

Modeling to find out the daily increase in case of coronavirus has become a hot topic lately. There are thousands of cases of coronavirus increased in Indonesia and the fluctuating value from day to day. Therefore, one of the techniques often used to solve predictive problems is modeling the Covid-19 case using machine learning techniques [4][5][6]. Research conducted by Mandayam et al. [4] and Gambhir et al. [5] using supervised machine learning techniques with the Regression method. Then research by Jarndal et al. [6] using Artificial Neural Network (ANN) method. The error rate of prediction model is usually assessed by Mean Squared Error (MSE), Root Mean Squared Error (RMSE) or Mean Absolute Error (MAE) [7]. The ANN and Regression methods still produce thousands of MSE values due to data fluctuation for the increased case of thousands of domains [4][6]. The data on the thousands increase of fluctuating Covid-19 needs to use a new predictive modeling process approach.

Artificial Intelligence can use the deep learning technique approach with the Long Short Term Memory (LSTM) method. Hochreiter and Schmidhuber proposed the LSTM method in 1997, that widely used to solve forecasting or prediction in various cases recently. The application of the LSTM method for prediction cases includes stock price prediction [8], mechanical state prediction [9], and water quality prediction [10]. The LSTM method for this prediction resulted in a value of either the MAE, MSE, or RMSE below 0.1. The error value close to zero indicates that the predicted number is getting closer to the actual result. Besides, this method is also the best model for dynamic data, which suddenly changes drastically in various cases [11]. Data on the addition of corona cases in Indonesia fluctuate in the number of thousands every day. Therefore, based on the LSTM method's success in minimizing the error value in other cases and this method is suitable for dynamic data, we propose the LSTM method to predict the Covid-19 case in Indonesia.

DOI: 10.34148/teknika.v10i1.328

This paper's contribution is related to the hyperparameter of the LSTM method used to predict Covid-19 data in Indonesia. The data used in this study came from Kawal Covid Indonesia [12]. The data processed in this study used daily increase cases. It is expected that the results of this study will provide a more accurate prediction representation using the LSTM method with an error value close to zero.

II. RESEARCH METHODOLOGY

Prediction of coronavirus cases in Indonesia begins with data acquisition. This study focuses on data on the daily increase cases. Before the data is processed into the next process, the data needs to be pre-processed because the Covid-19 cases fluctuate in the hundreds or even thousands. The pre-processing data results are carried out by making a model using the LSTM method. The model that is formed will be evaluated using MSE, RMSE, and MAE. The best results will be used to predict the daily increase cases. The flowchart of the coronavirus case prediction system is shown in Figure 1.

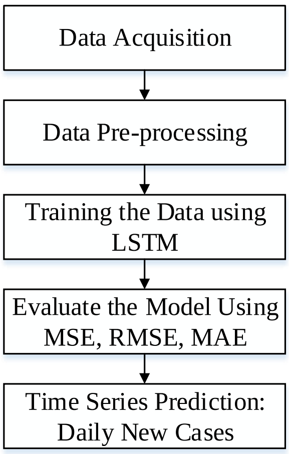

Data Acquisition
Data Pre-processing
Training the Data using LSTM
Evaluate the Model Using MSE, RMSE, MAE
Time Series Prediction: Daily New Cases

Figure 1. The Flowchart of Coronavirus Cases Prediction

A. Data Acquisition

Table 1. Time Series Data of Covid-19 Cases in Indonesia

Date	New Case
2020-03-11	7
2020-03-12	7
2020-03-13	35
2020-03-14	27
…	…
2021-01-01	8072
2021-01-02	7203
2021-01-03	6877
2021-01-04	6753

This study uses a dataset of daily Covid-19 cases obtained from the website of kawalcovid19.id in the spreadsheet document: https://docs.google.com/spreadsheets/d/1ma1T9hWbec1pXlwZ89WakRk-OfVUQZsOCFl4FwZxzVw/htmlview, which is an information portal about Covid-19 in Indonesia [12]. The data to be analyzed in this study uses data on new positive cases. Covid-19 used in this study is daily time-series data from March 11, 2020 to January 4, 2021. Time series data is a series of data sets ordered by time sequence [11]. Table 1 shows the Covid-19 data used in this study. The total data used in this study was 300 days.

B. Data Pre-processing

Data pre-processing is done before the data is processed using deep learning techniques of the LSTM method. Covid-19 data is divided into data for training and data for testing. The distribution of training data and testing for prediction is 80% training data and 20% testing data. The training data is normalized using the MinMax method so that data that is too large will not affect the modeling process. The MinMax method will convert data from zero to one. The MinMax scaler formula is shown in equation 1. Table 2 shows the results of data normalization.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where:
x' : normalized value
x : actual value
x_{min} : the minimum value of actual data
x_{max} : the maximum value of actual data

As an example of calculating on January 1, 2021, using equation (1).

$$x' = \frac{8072 - 7}{8369 - 7} = 0.96448218$$

Table 2. Covid-19 Data Normalization Results

Date	New Case
2020-03-11	0.0
2020-03-12	0.0
2020-03-13	0.00334848
2020-03-14	0.00239177
…	…
2021-01-01	0.96448218
2021-01-02	0.86055967
2021-01-03	0.82157379
2021-01-04	0.8067448

C. Training the Data Using LSTM

Long Short Term Memory (LSTM) is a model of deep learning derived from the development of the Recurrent Neural Network (RNN), which is also used to predict time series data [13]. The RNN can record previously used information. However, RNN has a weakness in vanishing-gradient, so that the prediction system for the Covid-19 case in Indonesia uses the LSTM method.

LSTM has the advantage of being able to recall long-term sequences (data measures), which are not easily achieved when using conventional methods. LSTM can use large data sizes and can use all the information as input to the system. LSTM has an architecture consisting of input, output, and hidden layers. The hidden layer consists of memory cells. One memory cell has three gates, namely the input gate, forget gate, and output gate [14]. The input of each LSTM unit is the previous output of the LSTM. Using time series data, the Covid-19 case's prediction now depends on the previous day.

The concept of the deep network in LSTM is shown in Figure 2. The implementation of LSTM in this research uses a Sequential approach. The first layer is the LSTM layer with the number of units. This research will test the number of neurons in the LSTM layer. A dropout is also applied to each layer to control overfitting. This research uses a dropout value of 0.25. The last layer is the output layer with a linear activation function. The loss parameter uses Mean Absolute Error (MAE) to find the minimum loss value in each iteration.

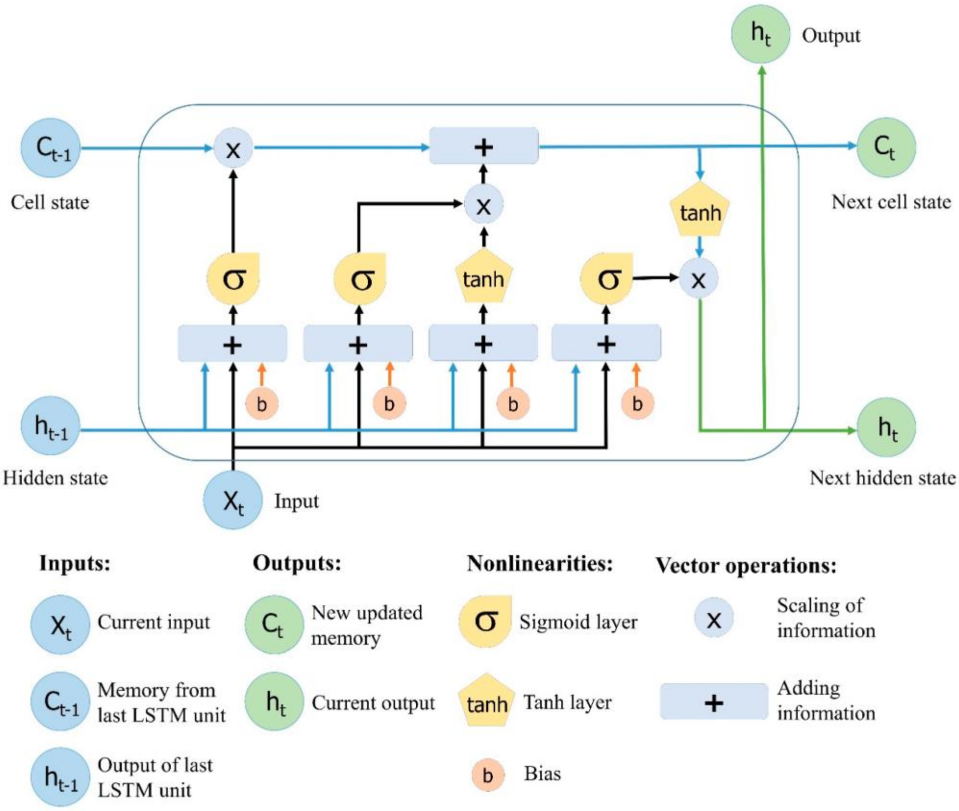
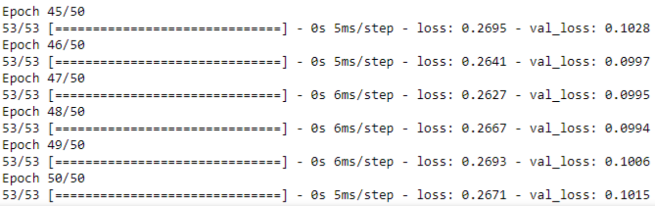


Figure 2. Illustration of the LSTM Method [15]

The result of pre-processing data, 240 data is used as training data, and 60 data for testing data. The sequence data is used as input to the LSTM. Figure 3 shows the process of training data using LSTM. In this study, validation data used 20% of the total training data used.



```
Epoch 45/50
53/53 [==============================] - 0s 5ms/step - loss: 0.2695 - val_loss: 0.1028
Epoch 46/50
53/53 [==============================] - 0s 5ms/step - loss: 0.2641 - val_loss: 0.0997
Epoch 47/50
53/53 [==============================] - 0s 6ms/step - loss: 0.2627 - val_loss: 0.0995
Epoch 48/50
53/53 [==============================] - 0s 6ms/step - loss: 0.2667 - val_loss: 0.0994
Epoch 49/50
53/53 [==============================] - 0s 6ms/step - loss: 0.2693 - val_loss: 0.1006
Epoch 50/50
53/53 [==============================] - 0s 5ms/step - loss: 0.2671 - val_loss: 0.1015
```

Figure 3. The Example Result of Training Data Using LSTM

D. Model Evaluation

The resulting model will be evaluated with matric such as the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) [16][17]. These three evaluation methods are used to measure the level of error prediction results. The smaller the value (close to 0), the prediction will be more accurate.

1. Mean Squared Error (MSE)

The MSE expresses the difference between the original and expected values derived by the squared average difference over the data set. By calculating the errors' middle squares, the error is now calculated in squared target units, which is the average squared difference between the expected values and the actual value. In Equation 2, the MSE formula is shown.

$$MSE(y, y') = \frac{1}{n}\sum_{i=1}^{n}(y_i - {y'}_i)^2 \quad (2)$$

where:

${y'}_i$: predicted value
y_i : actual value
n : the number of observation

2. Root Mean Squared Error (RMSE)

RMSE results from the square root of the MSE value that also measures the average magnitude of the error. It is usually used to evaluate matric and loss functions. The RMSE formula is shown in Equation 3.

$$RMSE(y, y') = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y'_i)^2} \tag{3}$$

where:
y'_i : predicted value
y_i : actual value
n : the number of observation

3. Mean Absolute Error (MAE)

MAE is almost the same as MSE. It is the mean of the absolute differences between forecast value and actual value over the test sample where all individual differences have equal weight. Without considering their course, MAE calculates the mean magnitude of the errors in a series of forecasts. The MAE formula is shown in Equation 4.

$$MAE(y, y') = \frac{1}{n}\sum_{i=1}^{n}|y_i - y'_i| \tag{4}$$

where:
y'_i : predicted value
y_i : actual value
n : the number of observation

III. RESULT AND DISCUSSION

A. Daily Covid-19 Cases Dataset

Daily increase cases data for Covid-19 is divided into two parts as training data and testing data. Figure 4 shows the division, the green line is training data, and the brown line is testing data. This study uses 80% training data and 20% testing data.

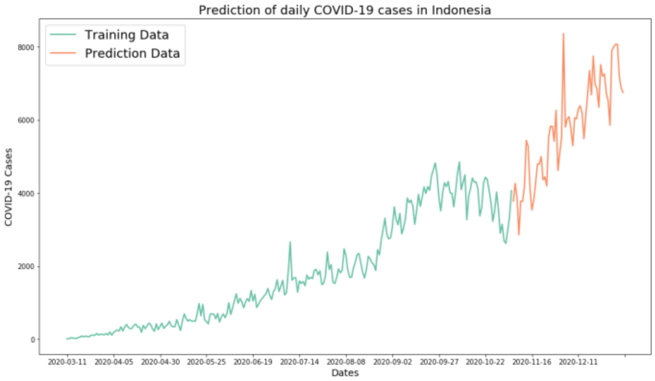


Figure 4. Data Split for Training and Testing

B. LSTM Model Configuration

The pre-processing dataset will be processed to the next stage, namely prediction processing using LSTM. Figure 5 shows a deep learning architecture using LSTM to form a prediction model for the Covid-19 case.

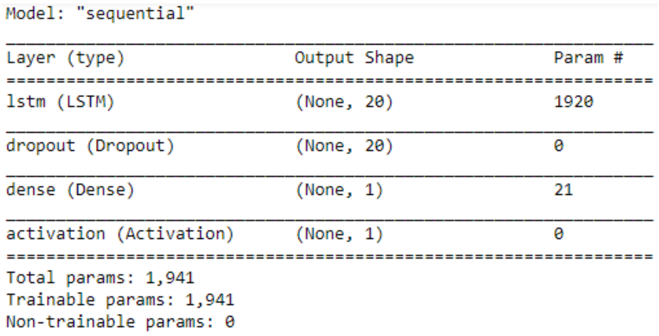

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, 20)                1920
_________________________________________________________________
dropout (Dropout)            (None, 20)                0
_________________________________________________________________
dense (Dense)                (None, 1)                 21
_________________________________________________________________
activation (Activation)      (None, 1)                 0
=================================================================
Total params: 1,941
Trainable params: 1,941
Non-trainable params: 0
```

Figure 5. Model Configuration

As shown in Figure 5, the LSTM model's configuration in this study is the LSTM layer, Dropout layer, Dense, and Linear Activation Function. This study uses the LSTM layer with 10, 25, 50, 100, and 200 units / hidden layers for the number of neurons. The Dropout layer functions to randomly select neurons that are dropped out or ignored during the training process. It results in other neurons having to enter to make predictions for neurons that have previously been removed. The dense layer is used to receive each neuron input from all the previous neurons to be connected, and the output will be generated on one neuron. Then, LSTM has a unique architecture to forget unnecessary information. This research uses the Linear Activation function. It takes the inputs and allows an output signal relative to the input, increased by each neuron's weights. In one way, as it requires multiple outputs, a linear function is superior to a phase function, not fair yes and no.

C. Model Evaluation

After the Covid-19 dataset is trained using the LSTM model, the next stage is testing the dataset to predict the addition of daily Covid-19 cases. In this study, we tested three hyperparameters for the prediction of daily Covid-19 cases in Indonesia. The three hyperparameters are the number of epochs, batch size, and the number of neurons. Each experiment was carried out five times because each experiment resulted in different numbers of MSE, RMSE, and MAE. The results we write down are the average of each trial.

1. The Experiment Number of Epochs

The first parameter in the experiment is the number of epochs. The LSTM model will use the number of neurons 25 and batch size 8. Table 3 shows the experimental results on the effect of the number of epochs.

Table 3. Experiment Results of Number of Epochs Effect

Epochs	MSE	RMSE	MAE
50	0.0354	0.188	0.1386
100	0.0364	0.1916	0.1454
200	0.0393	0.199	0.1542
400	0.0572	0.238	0.174
800	0.0776	0.277	0.197

DOI: 10.34148/teknika.v10i1.328

MSE, RMSE, and MAE data on the number of epoch experiments were obtained from an average of 5 trials for each epoch. Based on Table 3, the higher the number of epochs, the higher the error value. The best result is achieved at the number of epochs of 50. The number of epochs of less than 50 is not a significant error value. In the next experiment, the number of epochs will be 50.

2. The Experiment of Batch Size

The batch size parameter controls how often the network weights change. The best results in the number of epoch experiments are at epoch 50. In the batch size experiment, 50 epochs will be used, and the number of neurons is 25. Experiments on the effect of batch size on the error value are shown in Table 4.

The batch size parameter is one of the critical factors in testing and training data. Table 4 shows the results of the larger batch size, resulting in a smaller error value. In batch sizes 16 and 32, the error value does not significantly change. Therefore, in the window size experiment, batch size 32 was selected.

Table 4. Experiment Results of Batch Size Effect

Batch Size	MSE	RMSE	MAE
1	0.0386	0.1966	0.143
2	0.0376	0.1942	0.1422
4	0.0366	0.1928	0.1434
8	0.0354	0.188	0.1386
16	0.0348	0.1872	0.1364
32	0.0342	0.1862	0.1336

3. The Experiment Number of Neurons

The number of neurons in the LSTM affects the training network process. In general, the greater the number of neurons makes the training process longer and the learning of structures more. However, a large number of learning processes can cause overfitting.

Table 5. Experiment Results of Number of Neurons Effect

Neurons	MSE	RMSE	MAE
10	0.0308	0.1758	0.13
25	0.0342	0.1858	0.1362
50	0.0346	0.187	0.1386
100	0.0356	0.1896	0.1424
200	0.036	0.1902	0.1444

The experiment on the number of neurons used the number of epoch 50 and batch size 32. Table 5 shows that the more neurons, the higher the error value. The number of epochs between 25 and 50 results in an error value that is not too different. In this experiment, the minimum error is obtained from the number of neurons 10.

D. Discussion

The model created produces varied MSE, RMSE, and MAE values. The most influential hyperparameters are the number of epochs and the number of neurons. The number of epochs that are more than 200 will make the error value increase or can cause overfitting.

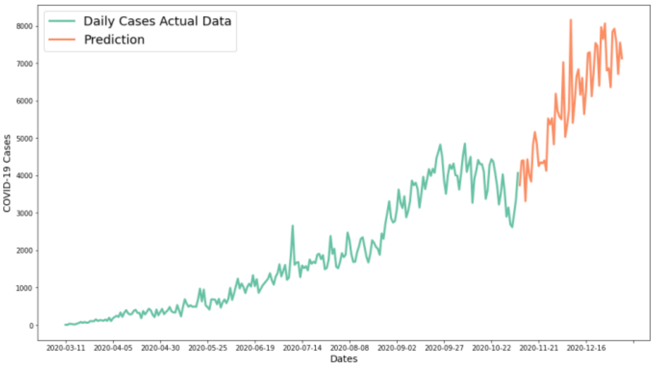


Figure 6. The Result of Daily Cases Prediction

Figure 6 shows the data for daily increase cases represented by a green line and the prediction results represented by a brown line. Then in Figure 7 shows the prediction results with actual data appearing coincided. It proves that the predicted value is close to the actual value.

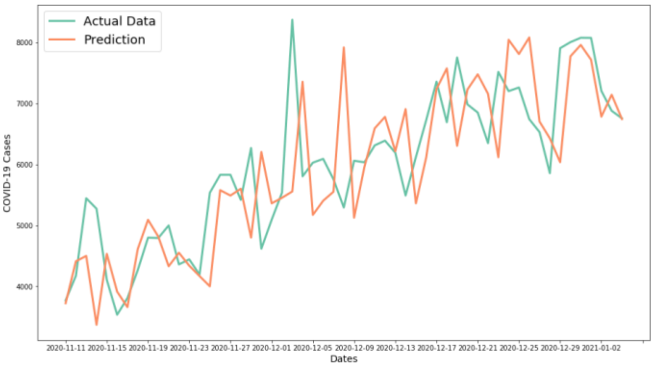


Figure 7. The Graph of Historical Daily Cases and Predicted Daily Cases

In this study, the best results were obtained with the LSTM model using the number of neurons 10, batch size 32, and epochs 50. The best results were obtained with an MSE error value of 0.03. The use of deep learning techniques results in smaller error values than using machine learning techniques. The value of MSE using machine learning techniques is still thousands [4][6]. Future research can modify the LSTM model to produce a minimum value of RMSE and MAE.

IV. CONCLUSION

This paper provides an overview of the prediction of the Covid-19 case in Indonesia. The number of Covid-19 cases in Indonesia has recently exploded, adding to the thousands per day. In the beginning, the addition of Covid-19 cases in Indonesia was only in units until recently, there have been thousands, and the data has fluctuated. Many previous studies used machine learning techniques, but the modeling results' error value was still thousands.

This study uses a deep learning approach to the LSTM method to build a model. The deep learning approach can minimize the error value in the resulting model. Based on several experiments, the minimum error values are MSE 0.0308, RMSE 0.1758, and MAE 0.13. Based on the prediction results, as shown in Figure 6, the addition of Corona cases in Indonesia is still increasing from day to day. Through these results, we suggest that the government issue stricter policies that corona case, mostly daily cases, can be pressed again, resulting in decreased significance. Future research can modify the LSTM model in order to produce a smaller error value.

ACKNOWLEDGEMENT

The author would like to thank the Kawal Covid Indonesia team for providing updated information on the Covid-19 case in Indonesia. This paper reflects only the authors' views and declares non-financial interest.

REFERENCES

[1] Pawar, M. (2020). The Global Impact of and Responses to the COVID-19 Pandemic. *The International Journal of Community and Social Development*, Vol. 2(2), pp. 111–120. https://doi.org/10.1177/2516602620938542.

[2] Pan, L., Mu, M., Yang, P., Sun, Y., Wang, R., Yan, J., Li, P., Hu, B., Wang, J., Hu, C., Jin, Y., Niu, X., Ping, R., Du, Y., Li, T., Xu, G., Hu, Q. & Tu, L. (2020). Clinical Characteristics of COVID-19 Patients With Digestive Symptoms in Hubei, China. *The American Journal of Gastroenterology*, Vol. 115(5), pp. 766–773. https://doi.org/10.14309/ajg.0000000000000620.

[3] Johns Hopkins University. (2021). *New Cases of COVID-19 In World Countries - Johns Hopkins Coronavirus Resource Center*. Johns Hopkins University. https://coronavirus.jhu.edu/data/new-cases.

[4] Mandayam, A.U., C, R. A., Siddesha, S. & Niranjan, S.K. (2020). Prediction of Covid-19 Pandemic Based on Regression. *Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, pp. 1–5.

[5] Gambhir, E., Jain, R., Gupta, A. & Tomer, U. (2020). Regression Analysis of COVID-19 Using Machine Learning Algorithms. *Proceedings of the International Conference on Smart Electronics and Communication (ICOSEC 2020)*, pp. 65–71. https://doi.org/10.1201/9781351073974.

[6] Jarndal, A., Husain, S., Zaatar, O., Gumaei, T.A. & Hamadeh, A. (2020). GPR and ANN based Prediction Models for COVID-19 Death Cases. *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, pp. 1–5. https://doi.org/10.1109/ccci49893.2020.9256564.

[7] Sulasikin, A., Nugraha, Y., Kanggrawan, J. & Suherman, A.L. (2020). Forecasting for a Data-Driven Policy Using Time Series Methods in Handling COVID-19 Pandemic in Jakarta. *2020 IEEE International Smart Cities Conference (ISC2)*, pp. 1–6. https://doi.org/10.2139/ssrn.3714105.

[8] Qian, F. & Chen, X. (2019). Stock Prediction Based on LSTM Under Different Stability. *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analytics, ICCCBDA 2019*, pp. 483–486. https://doi.org/10.1109/ICCCBDA.2019.8725709.

[9] Chen, Z., Liu, Y. & Liu, S. (2017). Mechanical State Prediction Based on LSTM Neural Network. *Chinese Control Conference (CCC)*, pp. 3876–3881. https://doi.org/10.23919/ChiCC.2017.8027963.

[10] Wang, Y., Zhou, J., Chen, K., Wang, Y. & Liu, L. (2017). Water Quality Prediction Method Based on LSTM Neural Network. *Proceedings of the 2017 12th International Conference on Intelligent Systems and Knowledge Engineering, ISKE 2017*, pp. 1–5. https://doi.org/10.1109/ISKE.2017.8258814.

[11] Bodapati, S., Bandarupally, H. & Trupthi, M. (2020). COVID-19 Time Series Forecasting of Daily Cases, Deaths Caused and Recovered Cases using Long Short Term Memory Networks. *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, pp. 525–530. https://doi.org/10.1109/iccca49541.2020.9250863.

[12] Kawal Covid 19. (2020). *Informasi Terkini COVID-19 di Indonesia*. Diakses dari: https://kawalcovid19.id/.

[13] Zhang, Y., Hutchinson, P., Lieven, N.A.J. & Nunez-Yanez, J. (2020). Remaining Useful Life Estimation Using Long Short-Term Memory Neural Networks and Deep Fusion. *IEEE Access*, Vol. 8, pp. 19033–19045. https://doi.org/10.1109/ACCESS.2020.2966827.

[14] Vinayakumar, R., Soman, K.P. & Poornachandran, P. (2017). Long Short-Term Memory Based Operation Log Anomaly Detection. *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) 2017*, pp. 236–242. https://doi.org/10.1109/ICACCI.2017.8125846.

[15] Le, X.H., Ho, H.V., Lee, G. & Jung, S. (2019). Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting. *Water (Switzerland)*, Vol. 11(7). https://doi.org/10.3390/w11071387.

[16] 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*, Vol. 13(7), pp. 914–919. https://doi.org/10.1016/j.jiph.2020.06.001.

[17] Rustam, F., Reshi, A.A., Mehmood, A., Ullah, S., On, B.W., Aslam, W. & Choi, G.S. (2020). COVID-19 Future Forecasting Using Supervised Machine Learning Models. *IEEE Access*, Vol. 8, pp. 101489–101499. https://doi.org/10.1109/ACCESS.2020.2997311.