Iraqi Journal of Science, 2023, Vol. 64, No. 11, pp: 5887- 5895
DOI: 10.24996/ijs.2023.64.11.34

ISSN: 0067-2904

Machine Learning Approach for New COVID-19 Cases Using Recurrent Neural Networks and Long-Short Term Memory

Intan Nurma Yulita[1]*, David Ferdinand Imanuel Manurung [2], Ino Suryana[2]
[1]*Research Center for Artificial Intelligence and Big Data, Universitas Padjadjaran, Bandung, Indonesia*
[2]*Department of Computer Science, Universitas Padjadjaran, Bandung, Indonesia*

Received: 15/5/2022 Accepted: 28/11/2022 Published: 30/11/2023

Abstract

This research aims to predict new COVID-19 cases in Bandung, Indonesia. The system implemented two types of deep learning methods to predict this. They were the recurrent neural networks (RNN) and long-short-term memory (LSTM) algorithms. The data used in this study were the numbers of confirmed COVID-19 cases in Bandung from March 2020 to December 2020. Pre-processing of the data was carried out, namely data splitting and scaling, to get optimal results. During model training, the hyperparameter tuning stage was carried out on the sequence length and the number of layers. The results showed that RNN gave a better performance. The test used the RMSE, MAE, and R2 evaluation methods, with the best numbers being 0.66975075, 0.47075, 0.29616625, and 0.7644 on the test data.

Keywords: Prediction, COVID-19, Long-Short Term Memory, Recurrent Neural Networks,

1. Introduction

Coronavirus or severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2) is a virus that attacks the respiratory system. The disease caused by this viral infection is called COVID-19. Coronavirus can cause mild disorders of the respiratory system, severe lung infections, and even death. According to the official website of the West Java government, on 15 May, 2022, 106,028 people in Indonesia were confirmed to be infected with the coronavirus. It is not a small number because Indonesia is ranked 7th in the number of COVID-19 cases on the Asian continent.

Many preventive measures have been taken by the government and the public to stop the spread of COVID-19, such as the implementation of the large-scale social restriction system in Bandung from April 22, 2020, to May 3, 2020. The restriction itself is a regulation made by the government to prevent the transmission of COVID-19. There are many rules, such as provisions for when the people of Bandung may be outside their homes, provisions for operating hours for places of business, and others. Although it aims to reduce the spread of COVID-19 in the city, there are side effects that are felt by many people in the city of Bandung. Many businesses cannot operate during the restriction period, disrupting the community's economy. With this situation, one solution is to predict the number of COVID-19 cases in the future so that this information can help the government make a policy on

*Email: intan.nurma@unpad.ac.id

whether it will be enforced again or not. Forecasting the number of confirmed cases of COVID-19 can be done using machine learning, especially RNN and LSTM.

Machine learning is the ability of a machine or computer to learn something [1]. With artificial intelligence (AI) embedded in a machine or computer, the machine can process the given input and give the desired output. It itself is a part of AI, which is more specifically working with statistics and data patterns to learn patterns. Just like the learning process in humans, machines need to be given examples or teachings so that they can understand what process should be followed. One of the machine learning methods itself is an artificial neural network.

Recurrent neural networks (RNN) are one type of artificial neural network. RNN has the property of being able to present sequential or time-series data. The processed data will be influenced by the previous data instance, so that it is said to be able to remember historical data [2]. Thus, the prediction of the number of COVID-19 cases in the city of Bandung can utilize machine learning technology as much as possible. Long short-term memory (LSTM) is an evolution of the RNN architecture that aims to make accurate predictions of a variable, where the variable in this case is the number of COVID-19 numbers. From many previous research results, the LSTM model is able to provide better performance than traditional machine learning models such as ARIMA [3]. The difference between the use of deep learning and traditional machine learning is the ability of deep learning to perform feature extraction and feature selection automatically.

Research that specifically addresses the prediction of new cases of COVID-19 can be found in a number of machine-learning studies. Yulita et al. studied it for a province in Indonesia with traditional machine learning [11]. The use of deep learning has also been applied to this prediction [12-13]. However, as we know, the pattern of the spread of COVID-19 in each region is different, so different models are needed. This research utilizes deep learning for prediction in a city in Indonesia. Through existing prediction models, local governments can better anticipate this disease.

2. Method

A “time series” is a series or sequence of events or observations taken sequentially over time. There are data points that will be related to the fixed-time method. The method is the process of analyzing the relationship between the variables in the data and the time variable. However, time series are just historical data without any relationship to future data. Therefore, by using this data, it can be used to make a prediction of what will happen in the future by processing existing time series data [14]. This study analyzed time-series predictions for new cases of COVID-19 in Bandung, Indonesia. This research was conducted in stages that include data collection, pre-processing (which is divided into data splitting and data scaling), model creation, and training. Figure 1 shows the flow of the research carried out.

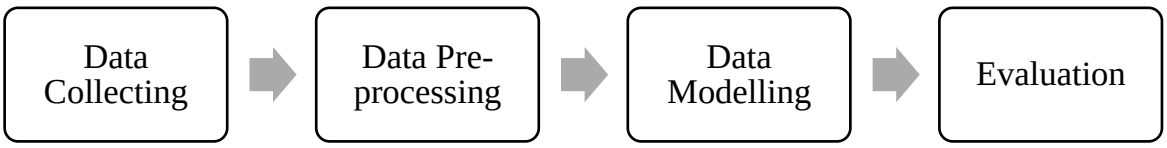

Figure 1: The system implementation

2.1 Data

The data used in this study is the number of new cases of COVID-19 in the city of Bandung from March 1, 2020, to December 31, 2020, which was taken from the official website of the *Coordination Center for COVID-19 Information and Coordination of West Java Province,* which can be accessed at the following link:

https://pikobar.jabarprov.go.id/data. The data contains two attributes: the date and the number of new COVID-19 cases in Bandung City on that date. As of December 31, 2020, there are 306 rows of data. The data will continue to increase over time, but in this study, the data used is limited to December 31, 2020. The data is described in Figure 2.

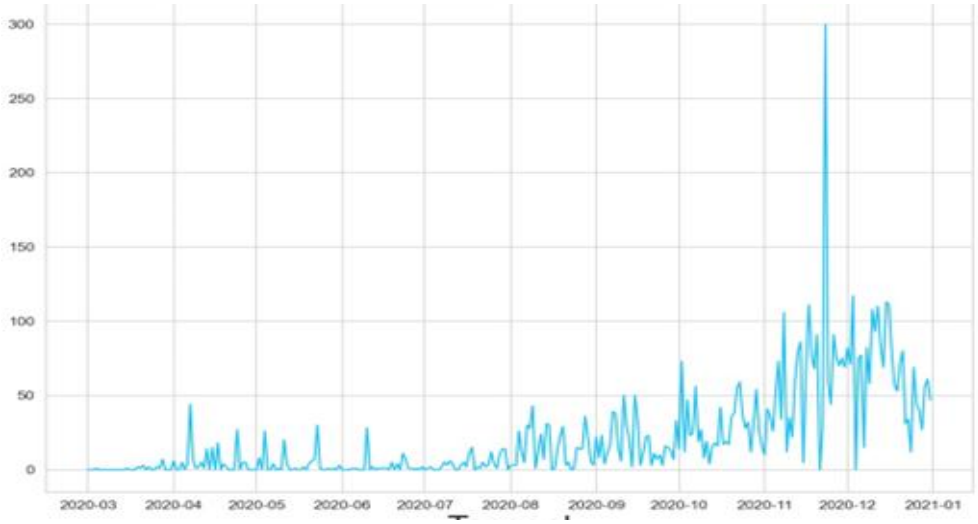

Figure 2: Number of Confirmed Cases of COVID-19 in Bandung

2.2 Data Pre-processing

Before the training process, the data needed to be processed first in order to produce better performance. The preprocessing stage in this research was the data splitting process and the data scaling process. Data splitting is a process where the overall data is divided into training data and test data [15]. The method used was holdout, which divided the entire data into 80% training data and 20% test data. The shared data will remain sequential because it includes time series data. Figure 3 is a visualization of the overall data division into training data and test data. The next preprocessing stage is the data scaling process using the min-max scaling method. The way min-max scaling works is that it adjusts the data within a certain range from a minimum to a maximum value. The range of values used is 0 to 1. Table 1 shows the data after normalization.

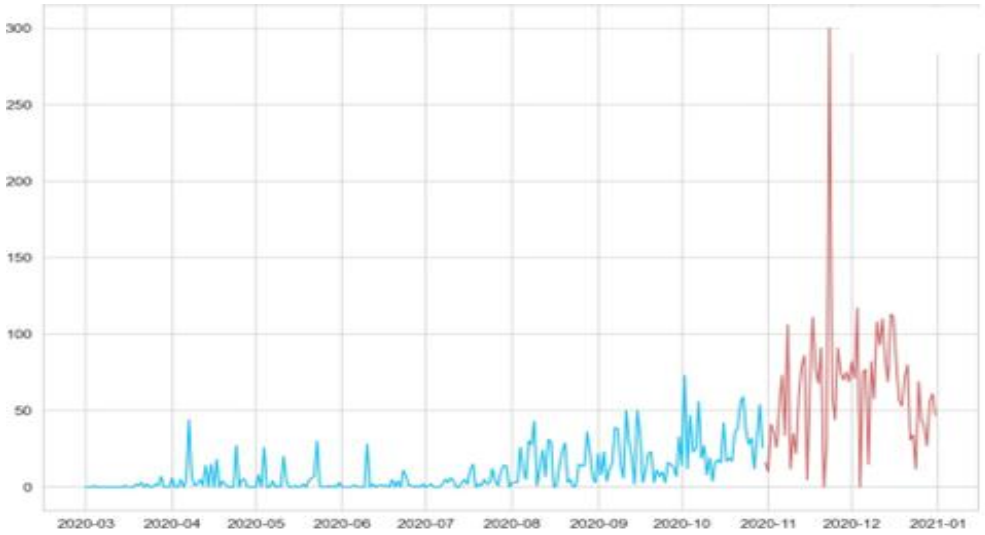

Figure 3: Splitting data: The blue and red lines show the train and test data, respectively

Table 1: Data after scaling

Date	Before scaling	After scaling
10/26/2020	32	0.438356
10/27/2020	12	0.164384
10/28/2020	32	0.438356
10/29/2020	54	0.739726
10/30/2020	26	0.356164

2.3 RNN and LSTM

RNN is one of the artificial neural network architectures where the output neurons will be reused and entered as input to the previous layer of neurons [16]. Thus, when processing data at time t, it will also have a weight value from time t–N [17]. The network can process errors or predictions from the past, which are described as output or hidden unit activities, for more precise and accurate future prediction calculations.

LSTM is an evolution of the RNN architecture that adds a memory cell that can store information for a long period of time [18]. LSTM can be a solution to the vanishing gradient problem owned by RNN, which causes RNN to fail to capture long-term dependencies, thereby reducing the accuracy of a calculation or prediction [19]. There are 3 different types of gate units used in LSTM: input gate, forget gate, and output gate. The input gate serves to determine whether an input will be added to the memory cell or not. The forget gate is useful for determining whether a memory from a previous time will be kept or forgotten. While the output gate is useful for determining how influential the memory in the cell state is on the results of calculations or predictions [20].

There is an activation function in the form of a sigmoid function in forget gates, where the result of the calculation is a Boolean value, namely 0 or 1. If the result is 1, then all data will be stored, and vice versa, if the result is 0, all data will be discarded. There are two activation functions in the input gates that are executed, namely the function to determine which value will be updated using the sigmoid function. The second is the tanh activation function to create a new vector value that will be stored in the memory cell. In cell gates, there is a function that will be executed, namely, a function that will replace the value in the previous memory cell with the new memory cell value, where this value is obtained by combining the values of the forget gate and input gate. In the cell gates, there are two functions that will be executed, namely the function to decide which part of the memory cell value will be issued; this function is in the form of a sigmoid function. The next function is a function to place a value in a memory cell with the tanh function. The results of the two gates are multiplied to produce the final output [21].

There are several hyperparameters in RNN and LSTM. Hyperparameters are variables in a model that will affect how the model works. The values will be determined before the model-training process. In this study, there will be several hyperparameters whose values will be changed to find the best value for each one; this process is called hyperparameter tuning. The hyperparameters that will be used in this study are the length of the sequence of inputs that will be entered into the model and the number of layers that will be used in each experiment. In each experiment, an epoch hyperparameter with a fixed value of 1000 epochs is also determined. One epoch is counted when all data has gone through one forward process and one backward process in the model training process. This study used only one variable, namely the number of confirmed cases of COVID-19. The hyperparameter tuning involved two variables, namely the length of the sequence and the number of LSTM layers. The sequence lengths used were 5, 7, 10, and 14. The number of LSTM layers used was 1, 2, 3, and 4. For each set, the hyperparameter was carried out five times, with 1000 epochs for each

experiment. It was done to find the average value of the evaluation and anticipate errors in the evaluation results.

2.4 Evaluation

There was an evaluation stage to re-examine the accuracy and performance results of machine learning. This study applied some evaluation methods to measure the level of accuracy of the forecasting method.

- The mean absolute error (MAE) is a method to calculate the average absolute error [22].

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}| \quad [1]$$

- Root Mean Squared Error (RMSE), an evaluation method that calculates the square of the error divided by the number of data and takes its root.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2} \quad [2]$$

- R^2 is an evaluation method that calculates the proportion of variance values described in the independent variables in the model. The result of the R^2 calculation gives a maximum value of 1. The closer the value is to 1, the better the evaluation value.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \quad [3]$$

Where:

$\hat{y}$: predicted value of y
$\bar{y}$: mean value of y

Results and Discussion

Hyperparameter tuning was done to find the hyperparameter that had the best performance. The evaluation scores RMSE, MAE, and R2 were determined through tests. Tables 2 and 3 show the hyperparameter tuning process on the RNN and LSTM models. The RNN model was better than the LSTM model. The best sequence length value for the RNN model was 14, and the number of layers was 1. However, if we look at the results of the data test on the RNN model with its best hyperparameters, the RNN model was overfitting. The model could not study the pattern in the data, so the results of the model trial on the RNN look like a straight line in Figure 4. LSTM showed a different pattern than RNN in Figure 5.

RMSE, MAE, and R2 values were best when the sequence length was 10, as shown in Table 3. The worst was 7. The effect of sequence length on model learning was very large. It affected the amount of data that entered the model to be trained. The data used as training data amounted to 80% of the total data, namely 244 pieces of data. A sequence length of 5 means the machine predicts a value by looking at 5 data points before the value is predicted. Likewise, when using sequence lengths of 7, 10, and 14, the fewer the sequences, the fewer patterns or information that can be learned by the machine, and the more difficult it is for the machine to predict future data trends. However, a sequence length that is too short will not ensure optimal machine learning. If the amount of data per sequence is too small, the machine cannot see the pattern or trend of the data in each sequence. Therefore, the optimal sequence length was 10 because it was not too short so the machine could learn the trend of the data. With limited data, stacked LSTM layers caused training and machine predictions to be more biased than their original values. It was also found that the most optimal number of LSTM layers is a layer that is not stacked, or one layer.

Table 2: The performance of RNN

Length	Num. of layer	Average		
		RMSE	MAE	R^2
5	1	0.866354	0.652400	-1.18474
	2	0.966152	0.757766	-1.70234
	3	0.893145	0.694523	-1.30877
	4	0.999926	0.798856	-1.99077
7	1	0.768695	0.572041	-0.71173
	2	0.765499	0.567380	-0.69470
	3	0.788569	0.593785	-0.79968
	4	0.792861	0.600480	-0.82167
10	1	0.838609	0.637579	-1.07606
	2	0.907931	0.710215	-1.41308
	3	0.799709	0.598825	-0.86309
	4	0.833802	0.627725	-1.01501
14	1	**0.669751**	**0.470750**	**-0.29617**
	2	0.675577	0.475165	-0.31888
	3	0.781793	0.585226	-0.77079
	4	0.691958	0.672827	-1.08878

Table 3: The performance of LSTM

Length	Num. of layer	Average		
		RMSE	MAE	R^2
5	1	0.828993	0.622829	-0.993082
	2	0884624	0.678285	-1.273695
	3	0.886306	0.676620	-1.310358
	4	0.949260	0.750865	-1.606780
7	1	0.817128	0.598559	-0.929908
	2	0.963380	0.770962	-1.690071
	3	0.979761	0.778987	-1.777524
	4	0.918683	0.720662	-1.438361
10	1	**0.764423**	**0.560741**	**-0.688390**
	2	0.778022	0.571493	-0.751919
	3	0.860426	0.660916	-1.139457
	4	0.917854	0.719790	-1.433958
14	1	0.769580	0.561351	-0.722330
	2	0.800520	0.587680	-0.854730
	3	0.916701	0.718579	-1.427850
	4	0.828993	0.622829	-0.993082

The best model was one that used a hyperparameter of sequence length 10 with a layer of LSTM. In other words, with RMSE 0.764423, MAE 0.560741, and R2 -0.688390, the model was not layered. Figure 5 is a visualization of the LSTM's results. There are two lines with different colors. The red line is the predicted data on the test data. It can be seen that the model could not follow the test data well. It was not caused by the failure of the model in the training process but by the characteristics of the initial dataset. In the test data, there was a lot of data that exceeded the peak data in the training data. During the model training process, the model never saw data with a value as high as the test data. Therefore, the model could not predict any data higher than the peak data in the training data. Although it could not predict spikes in data, the model seems to be able to follow up and down patterns in the data.

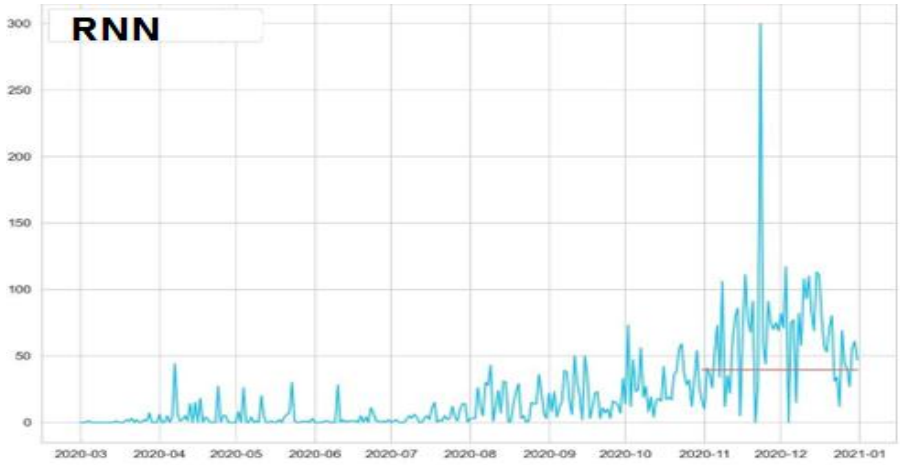

Figure 4: RNN's results: The blue and red lines show the actual dan predicted data, respectively.

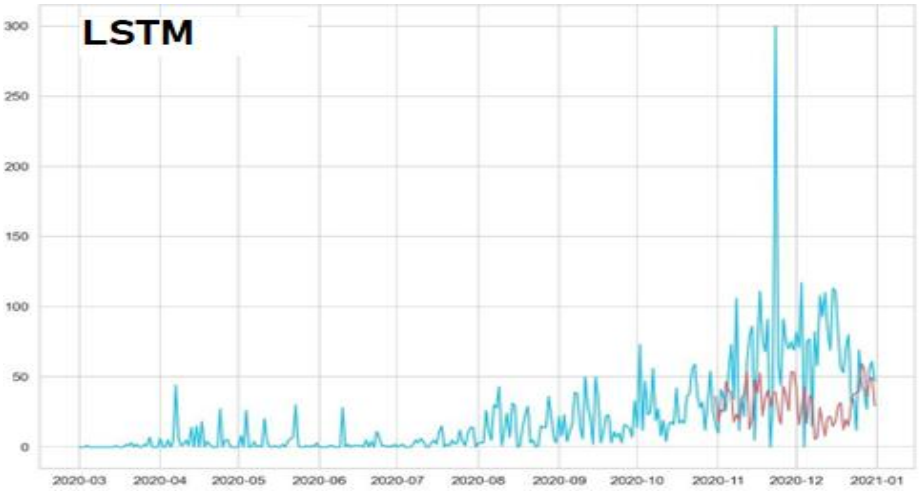

Figure 5: LSTM's results: The blue and red lines show the actual dan predicted data, respectively.

4. Conclusion

After conducting research, the research shows that the number of confirmed COVID-19 cases in the city of Bandung can be predicted using the RNN and LSTM models. But to be able to work optimally, it is necessary to do hyperparameter tuning on the hyperparameters that will be used in the model. In this study, two hyperparameters were selected: the length of the sequence and the number of layers. Both RNN and LSTM obtain optimal conditions by using a one layer LSTM model. According to the test results, the optimal RNN model built in this study has a performance with an RMSE value of 0.66975075, an MAE value of 0.47075, and an R2 value of -0.29616625. The best LSTM has an RMSE value of 0.764423, an MAE value of 0.560141, and an R2 value of -0.688390. It shows that RNN is better than LSTM. Our suggestions that can be considered for further development are related to data quality. The more data in the time series forecasting, the better the prediction of the model. The greater the amount of data, the more data trends will be formed, ranging from short-term to long-term data trends or even seasonal data trends. Therefore, the model's ability to remember information or data trends will be maximized. In this study, the data available was too little, but due to the urgency of the situation of the spread of COVID-19, a study was conducted to provide an overview and immediate results in handling cases of the spread of COVID-19 in

the city of Bandung. So, for further research, it is recommended to do research again so that it can help handle the spread of COVID-19 with the latest conditions. After carrying out the hyperparameter tuning process and getting the most optimal combination of hyperparameters, the model can be used to predict the number of new COVID-19 cases in the city of Bandung in the future.

5. Acknowledgements

We thanks to Rector Universitas Padjadjaran. Financial support was received from online data and a library research grant from Universitas Padjadjaran in 2020.

List of Abbreviations

Abbreviations	Definitions
COVID-19	Coronavirus Disease 2019
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Networks
ARIMA	Autoregressive Integrated Moving Average
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error

References

[1] M. M. Mijwil, "Implementation of Machine Learning Techniques for the Classification of Lung X-Ray Images Used to Detect COVID-19 in Humans," *Iraqi Journal of Science*, vol. 62, no. 6, pp. 2099-2109, 2020.

[2] M. C. Younis, "Evaluation of deep learning approaches for identification of different corona-virus species and time series prediction," *Computerized Medical Imaging and Graphics*, 90:101921, 2021.

[3] S.Siami-Namini, N. Tavakoli, and A. S. Namin, "A comparison of ARIMA and LSTM in forecasting time series," I*n 2018 17th IEEE international conference on machine learning and applications (ICMLA)*, 2018.

[4] M. U. D Khanday, Q. R. Khan, and S. T. Rabani, "Ensemble Approach for Detecting COVID-19 Propaganda on Online Social Network," *Iraqi Journal of Science*, vol. 63, no. 10, pp. 4488-449, 2022

[5] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, and A. E. Hassanien, "Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media," *Applied Soft Computing*, 97:106754, 2020.

[6] S Imran, S. M. Daudpota, Z. Kastrati, and R. Batra, "Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets," *IEEE Access*, vol. 8, pp. 181074-181090, 2020.

[7] M. E. Basiri, S. Nemati, M. Abdar, S. Asadi, and U. R Acharrya, "A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets," *Knowledge-Based Systems*, vol.228, 107242, 2021

[8] T. Liu, E. Siegel, and D. Shen, "Deep Learning and Medical Image Analysis for COVID-19 Diagnosis and Prediction," *Annual Review of Biomedical Engineering*, vol. 24, 2022.

[9] H. S. Alghamdi, G. Amoudi, S. Elhag, K. Saeedi, and J Nasser, "Deep learning approaches for detecting COVID-19 from chest X-ray images: A survey," *IEEE Access*, vol. 9, pp. 20235-20254, 2021.

[10] R. Jain, M. Gupta, S. Taneja, and D. J. Hemanth, "Deep learning based detection and analysis of COVID-19 on chest X-ray images," *Applied Intelligence*, vol. 51, no. 3, pp. 1690-1700, 2021

[11] Yulita, I. N., Abdullah, A. S., Helen, A., Hadi, S., Sholahuddin, A., & Rejito, J. "Comparison multi-layer perceptron and linear regression for time series prediction of novel Coronavirus covid-19 data in West Java," *In Journal of Physics: Conference Series*, vol. 1722, no. 1, p. 012021), 2021.

[12] N. Ayoobi, et al. "Time series forecasting of new cases and new deaths rate for COVID-19 using deep learning methods," *Results in Physics*, vol. 27, 104495, 2021.

[13] P. Arora, H. Kumar, and B. K Panigrahi. "Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India," *Chaos, Solitons & Fractals*, vol. 139, 110017, 2020.

[14] F. M. Khan, and R. Gupta, "ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India," *Journal of Safety Science and Resilience*, vol. 1, no. 1, pp. 12-18, 2018.

[15] K. De Souza, "Two cross-validation techniques to comprehensively characterize global horizontal irradiation regression models: Single data-splitting is insufficient," *Journal of Renewable and Sustainable Energy*, vol. 11, no. 6, 063702, 2019.

[16] Tokgöz, and G. Ünal, "A RNN based time series approach for forecasting turkish electricity load*," In 2018 26th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4), 2018.

[17] W. Yu, I. Y. Kim, and C Mechefske, "Analysis of different RNN autoencoder variants for time series classification and machine prognostics," *Mechanical Systems and Signal Processing*, vol. 149, 107322, 2021.

[18] K. magulova, and A. P. James, "A survey on LSTM memristive neural network architectures and applications," *The European Physical Journal Special Topics*, vol. 228, no. 10, pp. 2313-2324, 2019.

[19] M. Alhussein, K. Aurangzeb, and S. I. Haider, "Hybrid CNN-LSTM model for short-term individual household load forecasting," *IEEE Access*, vol. 8, pp. 180544-180557, 2020.

[20] F. Shahid, A. Zameer, and M. Muneeb, "Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM," *Chaos, Solitons & Fractals*, vol. 140, 2020.

[21] Y. Yu, X. Si, C. Hu, and J Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235-1270, 2019.

[22] W. Wang, and L. Lu, "Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model," *In IOP conference series: materials science and engineering*, vol. 324, no. 1, p. 012049, 2018.