Daily rainfall modeling using Neural Network

Daily rainfall modeling using Neural Network

S D Permai[a,*], M Ohyver[b] and M K B M Aziz[c]

[a,b] Statistics Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480
[c] Centre for Mathematical Sciences, College of Computing & Applied Sciences, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Kuantan, Pahang, Malaysia

[*]Corresponding author: syarifah.permai@binus.ac.id

Abstract. In the early 2020, Indonesia experienced flooding in several areas. This disaster caused a lot of damage and losses. One of the causes of flooding in Indonesia is due to high rainfall. This was not anticipated beforehand so there was a flood. Therefore, research on rainfall in Indonesia is very important to anticipate floods. If it is predicted that rainfall is very high and conditions do not allow it to accommodate, the government can prepare watersheds so that rainwater can flow and not be trapped. In this research, the rainfall data were obtained from Meteorological, Climatological, and Geophysical Agency (BMKG Indonesia), then the analysis of rainfall data in Indonesia was carried out. There are several statistical methods that can be used. There are ARIMA and Neural Network. In this research, the results of ARIMA model are used as input variables in the Neural Network model. Then there are several numbers of hidden layer in the Neural Network model that are compared. The results of ARIMA model and Neural Network model showed that Neural Network model is better than ARIMA model, because the mean square error (MSE) value of Neural Network model is smaller than ARIMA model.

1. Introduction

Floods are events that have occurred in Jakarta for a long time, even before Indonesia's independence. Some actions were even taken to overcome it, but it seems that it has not been successful. The construction of several canals could not prevent the flood at that time. The Governor of DKI Jakarta, Mr Ali Sadikin, had worked with foreign parties to build reservoirs in the city and the construction of new waterways. However, major floods still occurred in early 1976. During the leadership of Mr Sutiyoso, Jakarta experienced severe floods that caused casualties. There are 80 people died, 320,000 were displaced, as well as substantial material losses. Major floods then occurred again in 2015 and 2020. For rainfall, Meteorological, Climatological and Geophysical Agency (BMKG) noted that the rainfall that occurred in early 2020 was the highest rainfall since 1996 [1].

According to the National Disaster Management Agency (BNPB), floods in Jakarta in early 2020 submerged 308 urban communities with a maximum water level reaching six meters. While the death toll reached 60 people, with a total of 92,621 refugees scattered in 189 points. Such a large flood is not new in Jakarta. Prior to this, there were at least five major floods at DKI Jakarta in 2002, 2007, 2013 and 2014. The impacts of this such as death tolls, the distribution of flood points to the number of refugees, then it can be called 2007 to be the worst flooding in Jakarta [2].

Meteorological, Climatological and Geophysical Agency (BMKG) Indonesia explained that the occurrence of climate change increases the risk and opportunities of extreme rainfall so that it triggers floods in Jakarta. In addition, there was a repetition of high rainfall for several years. The cause of flooding in Jakarta is not only a matter of extreme rainfall and meteorological phenomena. However, there are several other factors such as the large amount of water runoff from the upstream area, the reduction in reservoirs and lakes where flood waters are stored. In addition, the problem of narrowing and the river's shallowness due to sedimentation and full of garbage, tidal soaking due to sea level and ground subsidence factors that increase the risk of standing water [3].

The negative impact caused by the flood needs to be done in prevent flooding. The problem is that there is no rainfall prediction that can be used to anticipate flooding. One of the efforts that can be done is to forecast rainfall data. Because if high rainfall can be predicted, action can be taken to overcome it. Therefore, in this research rainfall forecasting is done in Indonesia. There are several time series analysis methods that can be used for forecasting. The most commonly used method is ARIMA. But along with the development of technology and science, several new methods have emerged, one of which is Machine Learning. Because of this rainfall forecasting is done using the Neural Network. This study aims to perform daily rainfall modelling in Indonesia using several methods, namely ARIMA and Neural Network. Then the results of rainfall forecasting using ARIMA and Neural Network will be compared.

2. Rainfall

Indonesia has two types of seasons, that are rainy season and dry season. The problem that occurs in Indonesia during the rainy season is the occurrence of floods in several locations. One of the reasons is due to high rainfall in several locations, in addition to high altitude conditions in the area. Rainfall measurement is a measurement of the amount of rainwater that fall on a flat surface at a certain period in millimeters (mm) with the assumption that the collected rainwater in the flat does not evaporate, does not sink in and does not flow [4]. The rainfall of one mm means that a flat area of one square meter is accommodated by one millimeter or one liter of rainwater [5].

3. Autoregressive Integrated Moving Average (ARIMA) Model

Autoregressive Integrated Moving Average (ARIMA) is one of the time series methods that can use to forecast the value for next period. ARIMA models also known as Box-Jenkins models [6]. The stationarity of a time series is related to its properties in time. The definition of stationarity or weal stationarity are as follows.

- The expected value of the time series does not depend on time.
- The autocovariance function defined as $Cov(y_t, y_{t+k})$ for any lag k is only a function of k is only a function of k and not time: that is $\gamma_y(k) = Cov(y_t, y_{t+k})$.

For a time-invariant and stable linear filter and a stationary input time series x_t with $\mu_x = E(x_t)$ and $\gamma_k(k)= Cov(x_t, x_{t+k})$, the output time series y_t is also a stationary time series

$$E(y_t) = \mu_y = \sum_{-\infty}^{\infty} \quad \psi_i \mu_x \tag{1}$$

$$Cov(y_t, y_{t+k}) = \gamma_y(k) = \sum_{i=-\infty}^{\infty} \quad \sum_{j=-\infty}^{\infty} \quad \psi_i \psi_j \gamma_k(i-j+k) \tag{2}$$

General model of ARIMA(p,d,q)×(P,D,Q,)^S as follows.

$$\phi_p(B)(1-B)^d \Phi_P(B^S)(1-B^S)^D Y_t = \theta_q(B)\Theta_q(B^S)a_t \tag{3}$$

where B is the backshift operator, p is the order of non-seasonal AR, d is the differencing of non-seasonal and q is the order of non-seasonal MA. While S is time span of repeating seasonal pattern, P is the order of seasonal AR, D is the differencing of seasonal and Q is the order of seasonal MA.

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \tag{4}$$

$$\Phi_P(B^S) = 1 - \Phi_1 B^S - \Phi_2 B^{2S} - \cdots - \Phi_p B^{PS} \tag{5}$$

$$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q \tag{6}$$

$$\Theta_Q(B^S) = 1 - \Theta_1 B^S - \Theta_2 B^{2S} - \cdots - \Theta_Q B^{QS} \tag{7}$$

where $\phi_p(B)$ and $\theta_q(B)$ are the non-seasonal component of ARIMA model. Then $\Phi_P(B^S)$ and $\Theta_Q(B^S)$ are the seasonal component of ARIMA model. $\phi_1, \dots, \phi_p$ and $\Phi_1, \dots, \Phi_p$ are the parameters of AR for non-seasonal and seasonal component that must be estimated. While $\theta_1, \dots, \theta_q$ and $\Theta_1, \dots, \Theta_Q$ are the parameters of MA for non-seasonal and seasonal component that must be estimated.

4. Artificial Neural Network

Artificial Neural Network (ANN) is a nonlinear data processing that is built from an interconnected or connected which called a neuron and there is a weight for each connection. ANN was applied for data pattern identification analysis or data classification [7]. In this research, the data used is time series data, then the input layer in the Neural Network model is the lag of the time series at the target output.

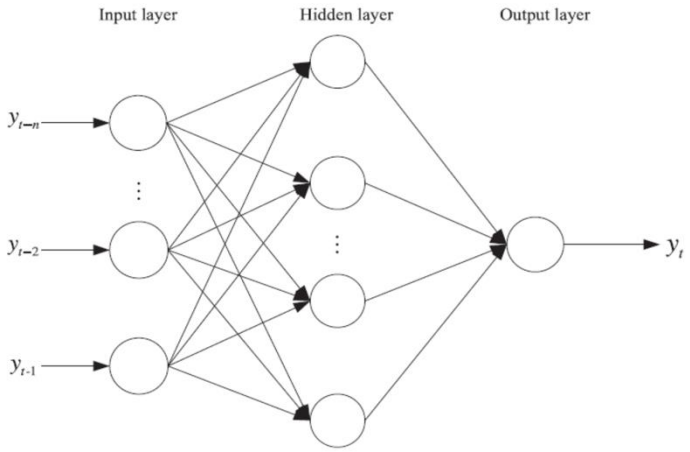

Figure 1. Single Hidden Layer of Neural Network

Figure 1 shows a single hidden layer of Neural Network. There are three layers namely input layer, hidden layer, and output layer [8].

$$I_j = \sum_{i=1}^{p} \beta_j + w_{ij}\, y_i \tag{8}$$

where p is the number of nodes in input layer, β_j is the threshold value of hidden layer, w_{ij} is the connection weight of input layer and hidden layer, and y_i is the lag of the time series in the input layer. Then using the activation function that has been determined in the hidden layer, the values in the hidden layer are obtained as follows.

$$Z_j = f(I_j) \tag{9}$$

The function for Z_j is called an activation function of hidden layer. There are linear and nonlinear activation functions. Some activation functions in ANN are follows [9].

- An identity or linear function
- The binary step function
- The binary sigmoid function or S-shaped curve
- The bipolar sigmoid function

Furthermore, the value of output layer can be calculate based on the activation function that has been determined in the output layer as follow [8].

$$H = \sum_{j=1}^{h} \alpha + v_j\, Z_j \tag{10}$$

$$y_t = f(H) \tag{11}$$

where h is the number of nodes in hidden layer, α is the threshold value of output layer, v_j is the connection weight of hidden layer and output layer, and y_t is the forecast value of point *t*. To get the thresholds and weights in the Neural Network model, the backpropagation algorithm is used to update these values. Backpropagation architecture is presented in the figure 2.

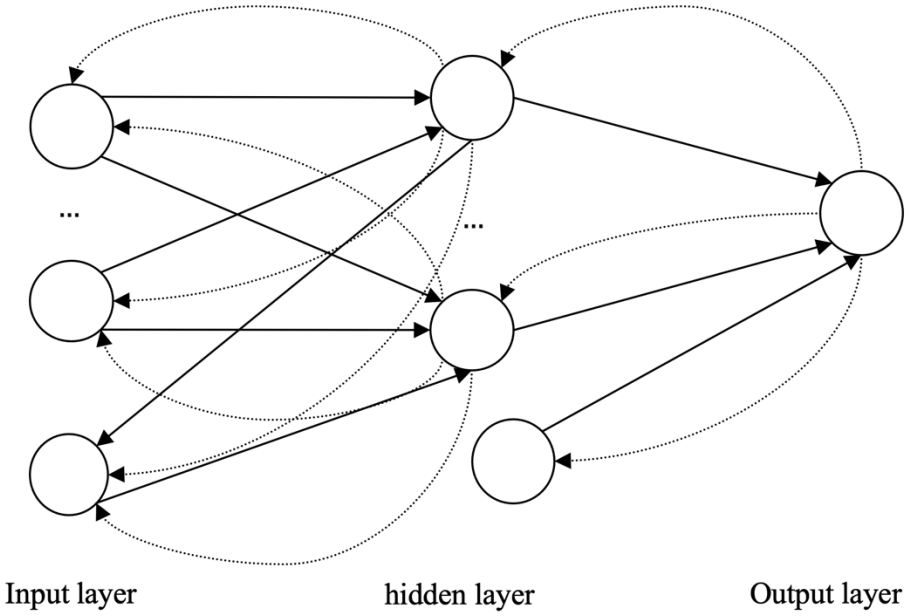

Figure 2. Backpropagation Architecture

In the backpropagation algorithm there are forward and backward phases. The straight arrow line in figure 2 shows the forward phase, while the dashed arrow line shows the backward phase. In the forward phase, weights are calculated starting from the input layer to the output layer using the specified activation function. Then the error value is obtained which is the difference between the forecast value with the target value. While in the backward phase, the error is propagated backward starting from the output layer to the input layer, then get new weights that minimize the error [9].

5. Criteria for selecting the best model

In this research, ARIMA and NN are used to model rainfall. Based on the two models, the best model will be chosen to the most appropriate model. In determining the selection of the best model, several criteria are needed. There are several criteria for selecting the best model including AIC, RMSE and MAE. AIC (Akaike Information Criteria) is one of the methods that used to select the best regression model found by Akaike and Schwarz [10]. According to Wei (2018), calculations for AIC can be calculated with the following formula [11].
follows.

$$\text{AIC} = \text{n ln } |\sum \text{p}| + 2\ \text{pm}^2 \tag{12}$$

n : number of observations
$|\sum p|$: determinant value of variance covariance matrix
p : number of input variables
m : number of output variables

According to Enders (2004), it is necessary to determine the appropriate number of lag dependent variables [12]. In determining the optimal lag there are several fairly efficient ways, such as

Akaike Information Criterion (AIC) and Schwarz 'Bayesian Information Criterion (SBC). The best model is the one that gives the smallest residual or error rate.

Root Mean Square Error (RMSE) is a popular method for assessing machine learning algorithms, including algorithms that are far more sophisticated than linear regression [13]. The RMSE value is used to distinguish the performance of the model in the trial period and the validation period [14]. Besides that, the Mean Absolute Error (MAE) can also be used to choose the best model. The MAE is the average of the absolute values of the difference between the forecast value and the observed value [15]. The RMSE and MAE value defined by the following formula.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} (y_t - \hat{y}_t)^2} \quad MAE = \frac{1}{n}\sum_{t=1}^{n} |y_t - \hat{y}_t| \tag{13}$$

where

n : number of observations
y_t : observed value
$\hat{y}_t$: forecasting value (14)

6. Data and Methodology

The data that used in this research were obtained from the Climatology, Meteorology and Geophysics Agency (BMKG) Indonesia [16]. The variable that be used in this research is rainfall data at Region I, which is the western region of Indonesia. The rainfall data used is the daily rainfall from January 2008 to December 2019. This data is time series data, then the analysis carried out is time series analysis. The stages of analysis to achieve the research objectives are as follows.

1. Estimating the parameters of rainfall model using ARIMA method with the following steps
   a. Identifying patterns on rainfall data
   b. Identifying ARIMA models
   c. Estimating the parameters of the ARIMA model
   d. Test the significance of parameters in the ARIMA model
   e. If the parameters are not significant, then back to step c
   f. Test the assumptions of the ARIMA model, which are the residuals have to white noise and normally distributed.
2. Estimation the parameters of rainfall model using Neural Network with Backpropagation algorithm, this is the following steps.
   a. Determine Feed Forward Neural Network (FFNN) architecture on rainfall data.
   b. Determine input for Neural Network architecture based on the best ARIMA model.
   c. Set the maximum epoch, target error and learning rate (α).
   d. Initialization parameters (thresholds and weights) in NN model.
   e. Make Network function on each node in hidden layer.
   f. Calculate the value of activation function at each node in hidden layer.
   g. Create a Network function on each node in output layer.
   h. Calculate the value of linear activation function at each node in output layer.
   i. Calculate the error values.
   j. Update parameters (thresholds and weights) using backpropagation algorithm.
   k. Repeat from step e and stop until the maximum number or target error is reached.
   l. Calculate the criteria of the best model selection for each model
   m. Choose the best NN model
3. Do forecasting of rainfall for testing data using the ARIMA and NN methods.
4. Determine the best model by comparing the AIC, MAE and RMSE values between ARIMA and NN model.

7. Results and Discussion

Before the modelling was performed to the rainfall dataset, data exploration was carried out using the autocorrelation function (ACF) plot and the partial autocorrelation function (PACF) plot. Figure 3 is the ACF and PACF plot of the rainfall data.

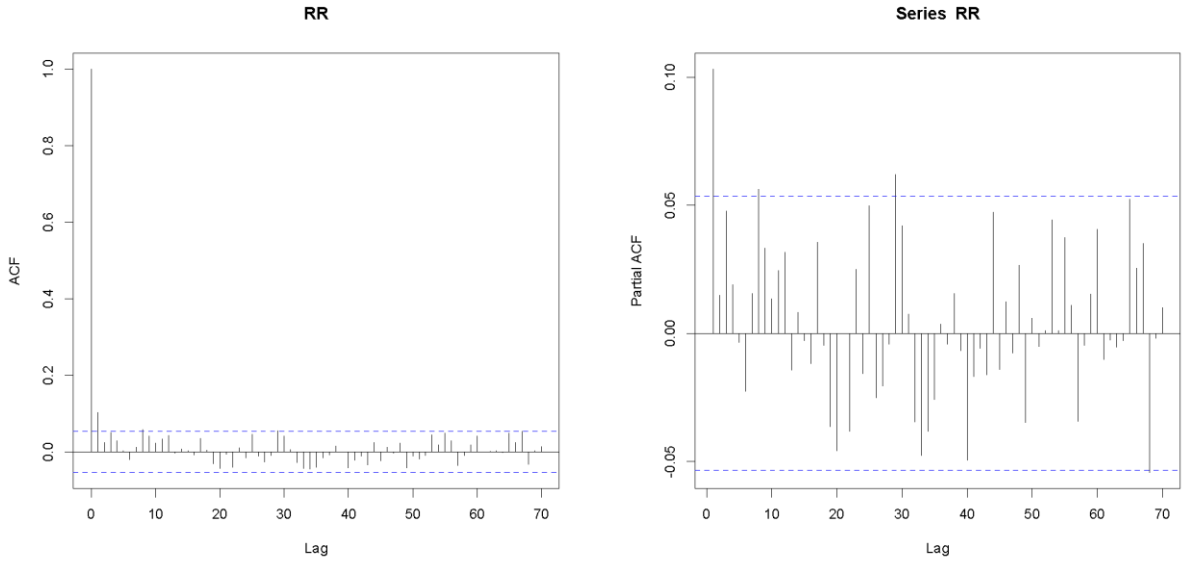

Figure 3. ACF and PACF plot

The identification results from the ACF and PACF plots showed that visually the data is stationary because the data is cut off at a certain lag. But to ensure stationarity in the rainfall data, the test was carried out using the Augmented Dickey-Fuller (ADF) test. The result of test statistics is -8.9025 and p-value is 0.01. The hypothesis of stationary test as follows.

H_0 : Data is not stationary

H_1 : Data is stationary

Because the p-value less than $\alpha = 10\%$ then H_0 is rejected. It means that the data is stationary. Then there is no need for transformations or differencing on the data. Furthermore, ARIMA modeling is carried out on rainfall data because the data is stationary. Then the ARIMA model does not use a difference (d = 0) because the data is stationary.

Based on the significant lag in the ACF and PACF plots, there are several ARIMA models as p and q value that can be tried. Because there is no indication of seasonal pattern then s = 0. Then ARIMA models only uses regular ARIMA which is ARIMA(p,d,q). There are several ARIMA models that have been applied on rainfall data that have significant parameters, including those presented in table 1.

Table 1. Comparison of ARIMA models.

Model	AIC	RMSE	MAE
ARIMA(0,0,1)	11048.73	14.85602	9.919381
ARIMA(1,0,0)	11048.3	14.85364	9.912884
ARIMA(1,0,1)	11048.88	14.84576	9.898771
ARIMA(2,0,1)	11048.24	14.83112	9.878218

If the orders of p and q increase, then the parameters are not significant in the model. Based on the result in table 1, it can be concluded that the best model is ARIMA(2,0,1) because this model has

the minimum AIC, RMSE and MAE values. Furthermore, the results of ARIMA(2,0,1) modelling is as follow.

Table 2. ARIMA(2,0,1).

Variable	Coefficient	Z test	p-value
AR1	0.960017	10.4527	<2e-16
AR2	-0.060654	-1.8875	0.0591
MA1	-0.862557	-9.8904	<2e-16
Intercept	8.135321	14.7300	<2e-16

Based on the parameter estimation results in table 2, the ARIMA model is obtained as follows.

$$\widehat{Y_t} = 8.135321 + 0.960017\, Y_{t-1} - 0.060654\, Y_{t-2} + 0.862557\, a_{t-1} + a_t \quad (15)$$

The hypothesis of testing the significance of the parameters.
H_0 : $\beta_i = 0$
H_1 : $\beta_i \neq 0$

Table 2 showed that the p-values of all parameters less than $\alpha = 10\%$. It means that all of parameters are significant in the ARIMA model. Furthermore, the residuals in the ARIMA model must meet the assumptions. Residuals have to follow the white noise process and residuals must be normally distributed. Ljung-box test is used to test the residuals follow white noise process and Kolmogrov Smirnov test is used to test the residuals following normally distribution.
The hypothesis of the Ljung-box test as follows.
H_0 : Residuals are white noise.
H_1 : Residuals are not white noise.

The result of Ljung-Box test showed that test statistics is 0.0013502 and p-value is 0.9707. Because the p-value greater than $\alpha = 10\%$ then do not reject H_0. It means that the residuals are white noise. Then the assumption of ARIMA model is fulfilled.
The hypothesis of the Kolmogorov Smirnov test as follows.
H_0 : Residuals are normally distributed.
H_1 : Residuals are not normally distributed.

The result of Kolmogorov Smirnov test showed that test statistics is 0.24235 and p-value = 2.2 x 10^{-16}. Because the p-value less than $\alpha = 10\%$ then reject H_0. It means that the residuals are not normally distributed. Then the assumption of ARIMA model is not fulfilled.

Since the ARIMA model did not fulfilled the assumption that residuals are normally distributed, then as an alternative to the ARIMA model, the Neural Network model can be used. In this research, the NN model uses a non-linear model, then the activation function used is logistic function or binary sigmoid function. Neural network modelling is very important in determining the number of inputs in the NN model. The input variables on NN model based on the significant lag in the ARIMA model. In this NN model, the input variables are Y_{t-1} and Y_{t-2}. This is based on the AR parameters in ARIMA model showing that the best ARIMA model is ARIMA(2,0,1). Then the next step is to determine the number of nodes in hidden layer. There are several numbers of nodes in hidden layer that have been tried in the Neural Network model to get the optimum number of nodes in the hidden layer. The comparison of Neural Network model for different number of nodes in the hidden layer can be seen in table 3.

Table 3. Comparison of Neural Network models.

Number of nodes in Hidden layer	AIC	RMSE
3 nodes	7240.006	14.80992
4 nodes	7246.006	14.80992
6 nodes	7238.144	14.700485

The others number of nodes in hidden layer does not converge then the parameters in the NN model are not obtained. Based on the optimum nodes in hidden layer that show in table 3, there are 6 nodes in hidden layer. Figure 4 show the best Neural Network model with 2 nodes in input layer, 6 nodes in hidden layer and 1 node in output layer. It can be written as NN(2,6,1).

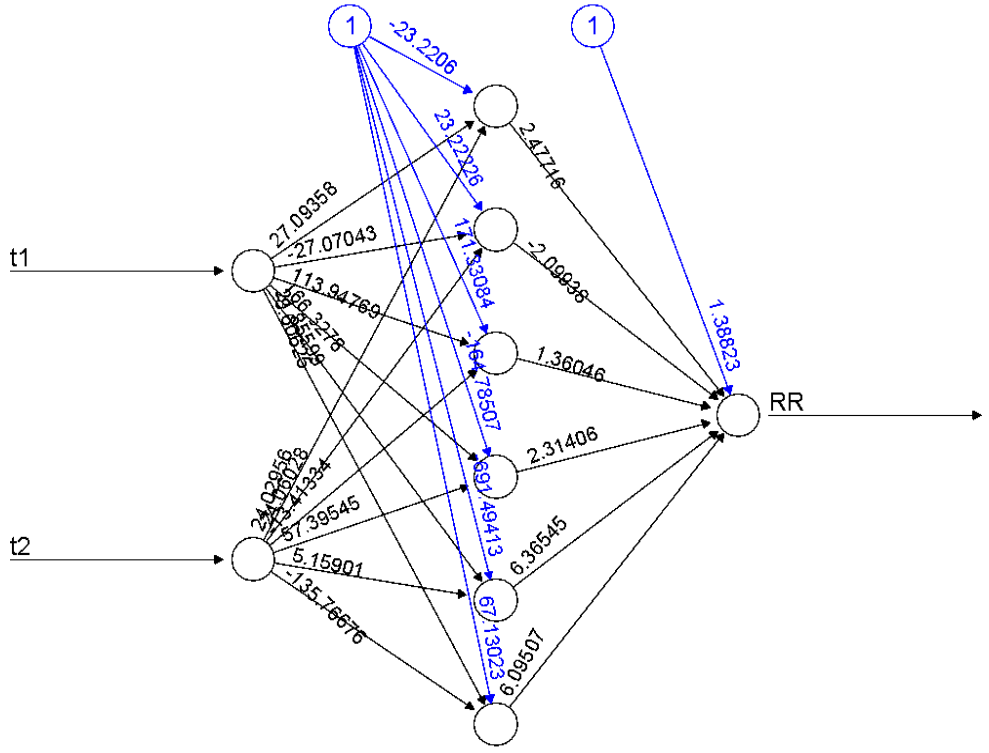

Figure 4. The result of neural network model with backpropagation algorithm

Table 4. Comparison of ARIMA and Neural Network model

Model	AIC	RMSE	MAE
ARIMA(2,0,1)	11048.24	14.83112	9.878218
NN(2,6,1)	7238.144	14.700485	9.8505714

Based on the results in table 4 showed that Neural Network model better than ARIMA model. It can be concluded that NN model better than ARIMA model because the AIC, RMSE and MAE value of NN model is smaller than ARIMA model. The AIC, RMSE and MAE value is an error value of the model, then the best model is the model that has a smaller value of AIC, RMSE and MAE.

8. Conclusion

There are several methods that can be used to analyze time series data. The Autoregressive Integrated Moving Average (ARIMA) model is one of the most frequently used methods for analyzing time series data. But this method can be use if the data is stationary and the model have some assumptions. The result of ARIMA model for rainfall data in this research showed that data is stationary but the assumption of residuals are normally distributed do not met in the ARIMA model. Therefore, the analysis is carried out using the neural network method as an alternative to the ARIMA model. The input variable of Neural Network model is determined from the significant lag in the ARIMA model. Then the input variables of NN model are Y_{t-1} and Y_{t-2}. Based on the experimental results, the number of nodes in the hidden layer concluded that the optimum number of nodes is 6 nodes in the hidden layer. Then in the NN model are determined 2 nodes in the input layer, 6 nodes in the hidden layer and 1 node in the output layer. The comparison between ARIMA model and NN model showed that NN model is better than ARIMA model. Because the RMSE and MAE value of NN model is smaller than ARIMA model. Based on this analysis, if we want to forecast the daily rainfall for today, we must consider the rainfall of yesterday and 2 days ago. There are some limitations in this research such as the number of nodes in the input layer, the activation function which only uses a logistic function, and the data is univariate time series. Then in the future research, we can use more nodes in the input layer, use other activation functions and try to use multivariate time series analysis.

Acknowledgments

This work is supported by Research and Technology Transfer Office, Bina Nusantara University as a part of Bina Nusantara University's International Research Grant entitled Rainfall Modeling to Prevent Flooding in Jakarta using Machine Learning Method with contract number: No.026/VR.RTT/IV/2020 and contract date: 6 April 2020.

References

[1] Putri, R. D. 2020. Mengungkap Musabab Banjir Besar di Jakarta. Tirto ID. [https://tirto.id/mengungkap-musabab-banjir-besar-jakarta-2020-eq85]

[2] Indonesia, CNN. 2020. Sejarah Banjir Besar Jakarta, Sejak Zaman VOC Hingga 2020.CNN Indonesia. [https://www.cnnindonesia.com/teknologi/20200102205129-199-462007/sejarah-banjir-besar-jakarta-sejak-zaman-voc-hingga-2020]

[3] Arnani, M. 2020. Penjelasan Lengkap Penyebab Banjir Jakarta, Curah Hujan Terekstrem hingga Sejarahnya. Kompas. [https://www.kompas.com/tren/read/2020/01/03/092500065/penjelasan-lengkap-penyebab-banjir-jakarta-curah-hujan-terekstrem-hingga?page=all]

[4] Mulyono, D. 2014. Analisis Karakteristik Curah Hujan di Wilayah Kabupaten Garut Selatan. *Jurnal Konstruksi*. Vol.12, No.1. E-ISSN: 2302-7312.

[5] Faradiba. 2018. Peramalan Curah Hujan dan Luas Serangan Organisme Pengganggu Tanaman di Kabupaten Bogor. *Jurnal Pro-Life*. Vol. 5, No. 3. E-ISSN: 2579-7557.

[6] Montgomery, D.C., Jennings, C.L., Kulahci, M. 2015. *Introduction to Time Series Analysis and Forecasting*. Second Edition. John Wiley & Sons. New Jersey. ISBN: 978-1-118-74511-3.

[7] Sivanandam, S.N., Sumathi, S., Deepa, S.N. 2006. *Introduction to Neural Networks Using Matlab 6.0*. Tata McGraw-Hill. New Delhi. ISBN: 9780070591127

[8] J. Singh and P. Tripathi. 2017. Time Series Forecasting Using Back Propagation Neural Network with ADE Algorithm. *International Journal of Engineering and Technical Research (IJETR)*, vol. 7, no. 5, pp. 19 - 23.

[9] Siang, J.J. 2005. *Jaringan Syaraf Tiruan dan Pemrogramannya Menggunakan Matlab*. Andi Offset. Yogyakarta.

[10] Grasa, A.A. 1989. *Econometrics Model Selection: A New Approach*. Springer. Netherlands. ISBN: 978-94-017-1358-0.

[11] Wei, W.W.S. 2018. *Time Series Analysis: Univariate and Multivariate Methods*. Second Edition. Pearson Addison Wesley. Boston. ISBN: 0-321-32216-9
[12] Enders, W. 2004. *Applied Econometric Time Series*. Second Edition. John Wiley & Sons. United States of America. ISBN: 0-471-23065-0.
[13] Conway, D., White, J.M. 2012. *Machine Learning for Hackers*. First Edition. O'Reilly Media, Inc. Cambridge.
[14] Gujarati, D.D., Porter, D.C. 2009. *Basic Econometrics*. Fifth Edition. McGraw-Hill/Irwin. New York.
[15] M. Małgorzata, M. Iwona, G. Magdalena and K. and Jaromir. 2018. Forecasting daily meteorological time series using ARIMA and regression models. *International Agrophysics*, vol. 32, no. 2, p. 253 – 264.
[16] BMKG. 2020. Data Online: Pusat Database - BMKG," BMKG Indonesia, 1 January 2020. [Online]. Available: https://dataonline.bmkg.go.id/. [Accessed 30 June 2020].