KHATULISTIWA SMART : JOURNAL OF ARTIFICIAL INTELLIGENCE
e-ISSN : 3089-9710
p-ISSN : XXXX-XXXX

Khatulistiwa : Journal of Artificial Intelligence
Vol. 3 No. 1 (2026) : 56 - 66
Available online at : https://ejournal.narasikhatulistiwa.org/index.php/kjarti

Rainfall Prediction in East Java Using Attention-Based LSTM

Beny Yusman [1]
[1] Universitas Hafshawaty Zainul Hasan, Indonesia
Beny.univer@gmail.com [1]

Doi

Submitted: 24 May 2026	Revision: 22 Jun 2026	Accepted: 25 Jun 2026	Published: 30 Jun 2026

Abstract Rainfall is one of the most influential meteorological parameters affecting agriculture, transportation, infrastructure, public health, and disaster mitigation. Accurate rainfall prediction is highly important, especially in East Java, one of the largest rice-producing regions in Indonesia. This study aims to develop a daily rainfall prediction model using the Attention-Based Long Short-Term Memory (LSTM) method. The dataset used in this study was obtained from the Indonesian Agency for Meteorology, Climatology, and Geophysics (BMKG), covering the period from January 1, 2000, to December 31, 2023, with a total of 8,764 daily records. The variables include average temperature, relative humidity, rainfall, sunshine duration, and average wind speed. The research stages include data collection, preprocessing, missing value handling, outlier replacement using the Z-Score method, correlation-based feature selection, Min-Max normalization, lagged feature construction, data splitting, Attention-Based LSTM model design, model training, model testing, and Streamlit application integration. The best model was obtained using a batch size of 16 and 100 epochs, achieving a validation MSE of 0.007038 and a validation RMSE of 0.083893. On testing data, the model achieved an MSE of 0.008076 and an RMSE of 0.089871. These results indicate that Attention-Based LSTM can model daily rainfall patterns effectively and has potential as a data-driven rainfall prediction support system.

Keywords *Attention-Based LSTM; Deep Learning; East Java; Rainfall Prediction; Time Series*

1. INTRODUCTION

Rainfall is one of the most influential meteorological parameters in daily life. High rainfall intensity can cause various impacts, such as infrastructure damage, transportation disruption, socio-economic losses, and increased risk of hydrometeorological disasters (Le et al., 2020). In Indonesia, rainfall prediction is increasingly important because the agricultural sector is highly dependent on climate conditions. East Java is one of the largest rice-producing regions in Indonesia, contributing 17.89% to national rice production in 2023 (Zulkipli, 2023). Therefore, accurate and timely rainfall prediction is needed to support decision-making in agriculture, disaster mitigation, and regional development planning. Rainfall prediction also plays an important role in reducing the impact of extreme weather events (Suleman & Shridevi, 2022), understanding climate change (Davenport & Diffenbaugh, 2021), evaluating ocean health (Xie et al., 2019), supporting weather forecasting (Singh et al., 2019), and preventing the spread of climate-related diseases (Rauf et al., 2019; Shahid et al., 2021). Although rainfall in a certain region cannot be determined with complete certainty, its intensity can be predicted using historical rainfall data (Setiawan & Barokah, 2022). The development of advanced computational models has made it possible to learn rainfall patterns more accurately through data-driven approaches (Kothai et al., 2023).

Various methods have been used for rainfall prediction, such as Backpropagation Neural Network (Sunardi et al., 2020), regression (Auliarahman, 2023), Support Vector Machine, Random Forest, and Linear Regression (Jalgaonkar & Kulkarni, 2021). Backpropagation has the ability to recognize complex patterns and tends to be stable in the learning process (Budiman et al., 2021). In addition, this method is widely used because it is relatively easy to understand in training multilayer neural networks (Sekhar & Meghana, 2020). However, Backpropagation has several limitations, such as slow convergence and the need for large amounts of training data (Falah et al., 2021; Lestari et al., 2023; Wahyudi & Pujiastuti, 2022).

Regression methods are advantageous because they are easy to understand and interpret (Secchin et al., 2020). However, regression generally assumes a linear relationship between independent and dependent variables. When the relationship between variables is non-linear, regression prediction performance may become less accurate (Zou et al., 2022). Meanwhile, Support Vector Machine and Random Forest have also been widely used in rainfall prediction. Support Vector Machine is known for its ability to find a global optimal solution and minimize overfitting risks in several classification and prediction cases (Sain et al., 2021). However, this method still faces challenges in parameter selection and risks of overfitting or underfitting (Raju et al., 2022). Random Forest performance can also be affected by very large and complex datasets (Nandini et al., 2022).

Most rainfall prediction studies have used relatively short data periods, either several years or a few decades. However, long-term historical rainfall data can provide more comprehensive information about climate fluctuations and long-term rainfall patterns. Therefore, the use of daily rainfall data from 2000 to 2023 in this study is expected to provide a stronger representation of rainfall dynamics in East Java. To address the limitations of previous methods, this study applies the Attention-Based LSTM method. Long Short-Term Memory (LSTM) has the ability to process time-series data and retain long-term information, while the attention mechanism helps the model focus on more important parts of the data (Han et al., 2024; Shali et al., 2020). Therefore, Attention-Based LSTM has the potential to improve rainfall prediction accuracy, especially for dynamic rainfall patterns with complex temporal dependencies.

Based on this background, this study aims to develop a daily rainfall prediction model in East Java using the Attention-Based LSTM method. The model was developed using BMKG data from 2000 to 2023 and evaluated using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). The results of this study are expected to contribute to the development of a more adaptive and accurate rainfall prediction model that can support disaster mitigation and sustainable development in East Java.

2. RESEARCH METHODS

2.1. Research Design

This study uses an experimental method with a deep learning approach to predict daily rainfall. The model used is Attention-Based Long Short-Term Memory (LSTM), which is an extension of the LSTM architecture integrated with an attention mechanism. The research stages include topic selection, problem identification, literature review, data collection, data preprocessing, model design, model training, model testing, result interpretation, and conclusion drawing.

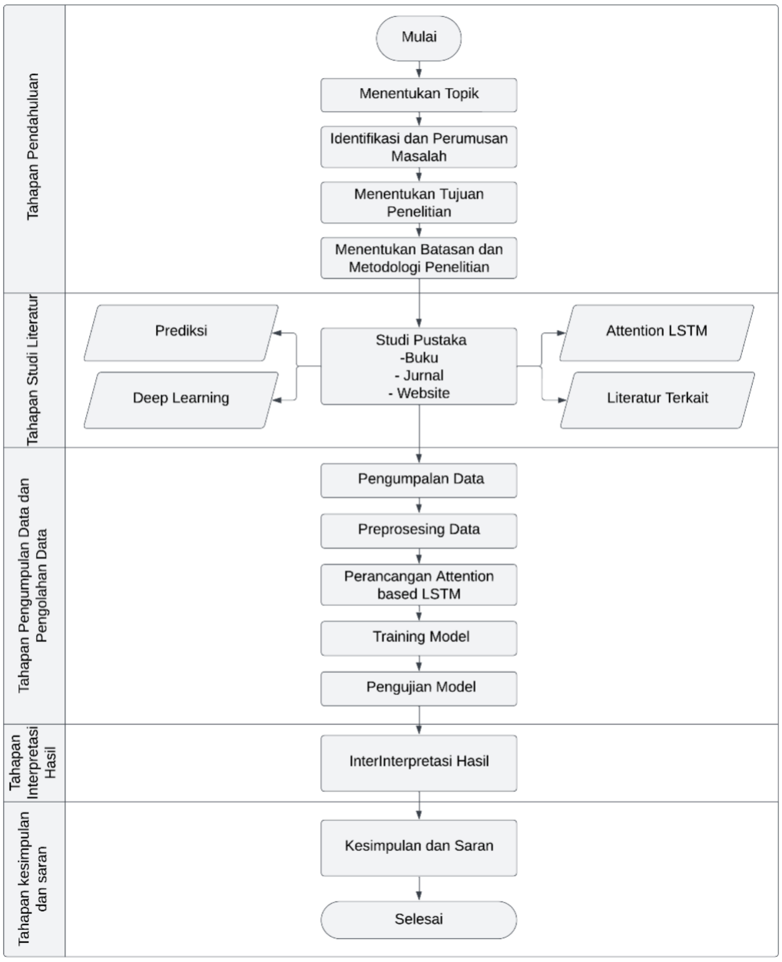

Figure 1. Research Framework

The research stages were arranged systematically so that data processing and model development could be carried out in a structured manner. The collected data first passed through preprocessing, then were transformed into a time-series format. After that, the data were used in the Attention-Based LSTM model training process and evaluated using MSE and RMSE.

2.2. Dataset

The data used in this study are secondary data obtained from the official website of the Indonesian Agency for Meteorology, Climatology, and Geophysics (BMKG) through http://dataonline.bmkg.go.id/data_iklim. The dataset consists of daily data from the East Java Climatology Station with station code 96943 from January 1, 2000, to December

31, 2023. The total number of data records used is 8,764. The variables used in this study are shown in Table 1.

Table 1. BMKG Dataset Variables

No	Variable	Description
1	Date	Date of data recording
2	Tavg (°C)	Daily average temperature
3	RH_avg (%)	Daily average relative humidity
4	RR (mm)	Daily rainfall
5	ss (hour)	Daily sunshine duration
6	ff_avg (m/s)	Daily average wind speed

The target variable in this study is daily rainfall or RR (mm). Meanwhile, the other variables are used as supporting features to help the model learn rainfall patterns.

2.3. Data Preprocessing

Data preprocessing was conducted to ensure that the dataset had good quality before entering the model training stage. The preprocessing process included date format conversion, missing value handling, outlier replacement, feature selection, normalization, data reshaping, and dataset splitting. The first step was converting the Date column into datetime format so that the data could be analyzed as time-based data. The next step was identifying missing values in each variable. The number of missing values before handling is shown in Table 2.

Table 2. Number of Missing Values Before Handling

Variable	Missing Values
Date	0
Tavg (°C)	33
RH_avg (%)	36
RR (mm)	217
ss (hour)	53
ff_avg (m/s)	437

Missing values were handled using forward fill and backward fill methods. Forward fill replaces missing values with previous values, while backward fill replaces missing values with subsequent values. After this process, all variables had zero missing values. The next step was outlier handling using the Z-Score method. Values with an absolute Z-Score greater than 3 were considered outliers and replaced with the median value. This method was used to prevent extreme values from causing excessive distortion during the model training process.

2.4. Feature Selection

Feature selection was conducted to determine the relationship between independent variables and the target variable RR (mm). Pearson correlation analysis was used to examine the strength of the relationship between variables. The correlation results are shown in Table 3.

Table 3. Feature Correlation Results Against Rainfall

Variable	Correlation Value Against RR (mm)
RH_avg (%)	0.326558
Tavg (°C)	0.040954
ss (hour)	-0.257047
ff_avg (m/s)	-0.181402

Based on the correlation results, RH_avg (%) has the highest positive relationship with rainfall. The variable Tavg (°C) is still used because it contributes as a supporting meteorological parameter. Meanwhile, ss (hour) and ff_avg (m/s) were removed because they had lower correlations with rainfall. Thus, the features used in modeling are Tavg (°C), RH_avg (%), and RR (mm).

2.5. Normalization and Data Reshaping

Normalization was performed using Min-Max Scaler so that all data values were within the range of 0 to 1. Normalization is necessary to ensure that each feature has a uniform scale and that no variable dominates the model learning process. An example of normalized data is shown in Table 4.

Table 4. Example of Normalization Results

Tavg (°C)	RH_avg (%)	RR (mm)
0.557143	0.325581	0.565517
0.385714	0.511628	0.000000
0.514286	0.534884	0.096552
0.571429	0.511628	0.068966
0.457143	0.534884	0.000000

After normalization, the data were transformed into lagged features using a timestep of 7 days. This means that the model uses information from the previous seven days to predict rainfall on the following day. The data reshaping results are shown in Table 5.

Table 5. Model Input Dimensions

Data	Dimension
X	(8758, 7, 3)
y	(8758,)

The dimension of X indicates that there are 8,758 samples, each consisting of 7 timesteps and 3 features. Meanwhile, y represents the rainfall target to be predicted.

2.6. Data Splitting

The data were divided into three parts: training data, validation data, and testing data. The splitting was performed sequentially without randomization to preserve the chronological structure of the time-series data. The data were divided into 80% training data, 10% validation data, and 10% testing data. The data splitting results are shown in Table 6.

Table 6. Dataset Splitting

Data	Number of Samples
X_train	7006
y_train	7006
X_val	876
y_val	876
X_test	876
y_test	876

2.7. Attention-Based LSTM Model

The model used in this study is Attention-Based LSTM. LSTM is a part of Recurrent Neural Network (RNN) that has the ability to store long-term information through gate mechanisms, namely forget gate, input gate, and output gate (Han et al., 2024). LSTM is effective for time-series data because it can capture temporal dependencies. The attention mechanism was added to help the model assign greater weight to timesteps considered more important in the prediction process. Attention allows the model to focus on more relevant information and reduce the influence of less important information (Han et al., 2023). The model architecture is shown in Table 7.

Table 7. Attention-Based LSTM Model Architecture

Layer	Output Shape	Parameters
Input	(None, 7, 3)	0
LSTM	(None, 7, 128)	67,584
Dropout	(None, 7, 128)	0
LSTM	(None, 7, 128)	131,584
Dropout	(None, 7, 128)	0
Attention	(None, 7, 128)	0
Flatten	(None, 896)	0
Dense	(None, 1)	897
Total Parameters		**200,065**

The model consists of two LSTM layers, two Dropout layers, one Attention layer, one Flatten layer, and one Dense output layer. The model has a total of 200,065 parameters, all of which are trainable.

2.8. Model Training and Evaluation

Model training was conducted using the Adam optimizer and the tanh activation function. Hyperparameter testing was performed using batch sizes of 4, 16, 32, 64, and 128, and epochs of 10, 20, 50, and 100. Evaluation was conducted using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). MSE is used to calculate the average squared error between actual values and predicted values. RMSE is used to calculate the square root of MSE, making the interpretation easier because it is on the same scale as the target variable.

3. RESULTS

3.1. Data Collection Results

The data used in this study are daily BMKG data from January 1, 2000, to December 31, 2023. The dataset consists of 8,764 rows with five main meteorological variables: average temperature, average relative humidity, rainfall, sunshine duration, and average wind speed. An example of the initial dataset is shown in Table 8.

Table 8. Example of BMKG Data

Date	Tavg (°C)	RH_avg (%)	RR (mm)	ss (hour)	ff_avg (m/s)
01/01/2000	24.1	69	82	5.4	2
02/01/2000	22.9	77	0	1.2	4
03/01/2000	23.8	78	14	4.1	3
04/01/2000	24.2	77	10	1.3	2
05/01/2000	23.4	78	8888	3.6	4

The initial dataset contained extreme values, such as 8888 in the RR (mm) variable, which was then handled as an outlier using the Z-Score method.

3.2. Data Preprocessing Results

The preprocessing results show that all missing values were successfully filled using the forward fill and backward fill methods. After that, outliers in several variables were replaced using the median value. One example of an outlier was the RR (mm) value of 8888, which had a Z-Score greater than 3 and was therefore replaced with the median value. After the feature selection process, the variables retained were Tavg (°C), RH_avg (%), and RR (mm). These variables were then normalized using Min-Max Scaler and transformed into lagged features with a 7-day timestep. The data were then divided into training, validation, and testing sets with 7,006, 876, and 876 samples, respectively.

3.3. Model Architecture Results

The Attention-Based LSTM model developed in this study has a total of 200,065 parameters. The model consists of two LSTM layers with 128 units each, two Dropout layers, one Attention layer, one Flatten layer, and one Dense layer to generate rainfall predictions. The Attention layer is used to improve the model's ability to select temporal information that is more relevant to rainfall prediction.

3.4. Training Results

Training was conducted using various combinations of batch sizes and epochs. The training and validation evaluation results are shown in Table 9.

Table 9. Training and Validation Metrics

Batch Size	Epochs	Training MSE	Training RMSE	Validation MSE	Validation RMSE
4	10	0.006500	0.080624	0.007491	0.086552
4	20	0.006253	0.079081	0.007338	0.085662
4	50	0.006028	0.077645	0.007181	0.084742
4	100	0.005905	0.076845	0.007165	0.084648
16	10	0.006545	0.080903	0.007590	0.087124

16	20	0.006189	0.078674	0.007339	0.085673
16	50	0.006018	0.077578	0.007185	0.084766
16	100	0.005845	0.076455	0.007038	0.083893
32	10	0.006482	0.080516	0.007526	0.086754
32	20	0.006367	0.079794	0.007470	0.086429
32	50	0.005998	0.077447	0.007146	0.084537
32	100	0.005888	0.076734	0.007089	0.084198
64	10	0.006526	0.080787	0.007528	0.086768
64	20	0.006452	0.080328	0.007472	0.086441
64	50	0.006068	0.077900	0.007214	0.084940
64	100	0.005951	0.077144	0.007110	0.084322
128	10	0.006520	0.080746	0.007508	0.086649
128	20	0.006468	0.080429	0.007577	0.087049
128	50	0.006039	0.077716	0.007165	0.084647
128	100	0.005922	0.076958	0.007054	0.083991

Based on Table 9, increasing the number of epochs generally reduced MSE and RMSE values in both training and validation data. The model with a batch size of 16 and 100 epochs achieved the best performance, with a validation MSE of 0.007038 and a validation RMSE of 0.083893. This result indicates that this configuration had better generalization capability compared to the other configurations.

3.5. Model Testing Results

After the best model was obtained, testing was conducted using testing data. The testing results show that the model achieved an MSE of 0.008076804690063 and an RMSE of 0.0898710447811919. These values indicate that the average prediction error of the model was relatively small after the data had been normalized.

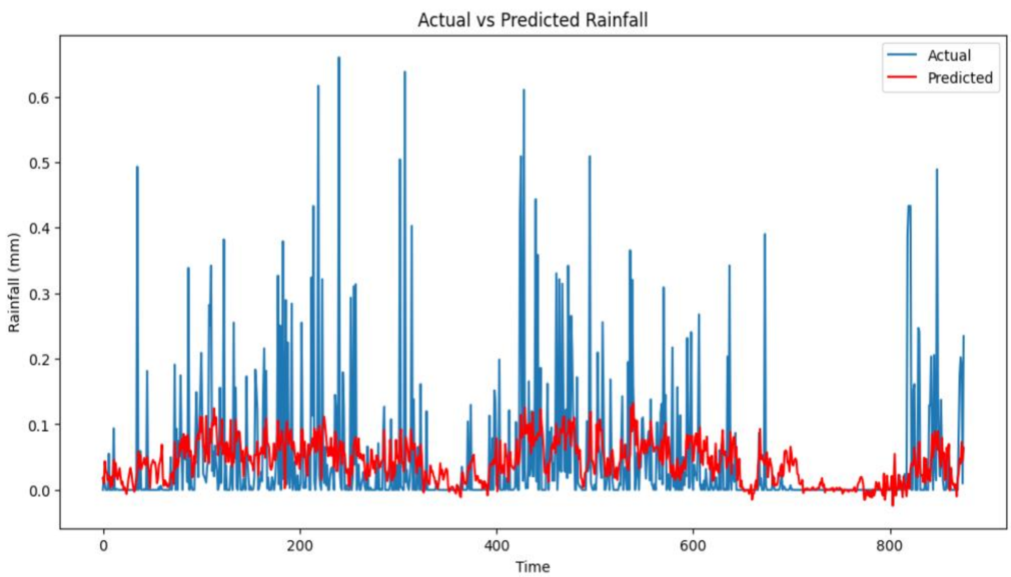

Figure 2. Actual vs Prediction Graph

The application provides data upload features, model selection, prediction processing, result visualization, and prediction result download options. The application testing results show that the predictions generated by the application are consistent with the model testing results in Google Colab. Therefore, the Streamlit implementation can make the rainfall prediction model more practical and easier to use.

3.6. Discussion

The results of this study show that Attention-Based LSTM can predict daily rainfall with good performance. The testing RMSE value of 0.089871 is slightly higher than the validation RMSE value of 0.083893. This difference is still within an acceptable range and indicates that the model maintains its generalization capability on unseen data. Compared with previous studies, the results of this study show competitive performance. The comparison of evaluation metrics is presented in Table 10.

Table 10. Comparison with Previous Studies

Author	Method	Evaluation Results
Lattifia et al. (2022)	Long Short-Term Memory	RMSE: 1.7444; MAPE: 1.9499%
Yusuf et al. (2022)	Multiple Linear Regression	MAE: 78.8695; RMSE: 95.1982
Patriardian et al. (2023)	ARMA	MAE: 175.372; RMSE: 246.647
This study	Attention-Based LSTM	MSE: 0.00807; RMSE: 0.08987

Based on Table 10, the Attention-Based LSTM method produced low MSE and RMSE values. However, metric comparisons across studies should be interpreted carefully because each study used different datasets, regions, data periods, normalization scales, and input variables. Nevertheless, the results of this study indicate that Attention-Based LSTM has strong potential for predicting daily rainfall in East Java.

4. CONCLUSIONS

Based on the results of this study, it can be concluded that the Attention-Based LSTM method can be used to predict daily rainfall in East Java. The model was developed using BMKG data from January 1, 2000, to December 31, 2023, with average temperature, relative humidity, and rainfall as the main features. The preprocessing process included missing value handling, outlier replacement using Z-Score, feature selection, Min-Max normalization, lagged feature construction with a 7-day timestep, and data splitting into training, validation, and testing sets. The training results show that the best configuration was obtained using a batch size of 16 and 100 epochs, with a validation MSE of 0.007038 and a validation RMSE of 0.083893. On testing data, the model achieved an MSE of 0.008076 and an RMSE of 0.089871. These results indicate that the model has a reasonably good ability to follow actual rainfall patterns, although several deviations still occur at certain points. The integration of the model into a Streamlit application shows that the model can be used more practically for rainfall prediction through a user interface. The application allows users to upload data, run preprocessing, perform prediction, display actual vs prediction graphs, and download prediction results. Thus, this study contributes to the development of a deep learning-based rainfall prediction system. This study still has limitations, particularly because the features used only include several meteorological variables. Future studies are recommended to add other variables, such as air pressure, wind direction, maximum and minimum temperature, climate indices, and spatial data. In addition, comparisons with other models such as GRU, CNN-LSTM, Transformer, or hybrid deep learning models can be conducted to improve prediction accuracy and model generalization.

5. REFERENCES

Auliarahman. (2023). Prediksi curah hujan menggunakan metode regresi. Jurnal Prediksi Cuaca dan Iklim, 5(2), 45–53.

Budiman, B., et al. (2021). Penerapan jaringan saraf tiruan backpropagation dalam pengenalan pola. Jurnal Teknologi Informasi dan Komputasi, 8(1), 12–20.

Davenport, F. V., & Diffenbaugh, N. S. (2021). Climate change and precipitation extremes. Climate Dynamics, 57(3), 789–802.

Falah, F., et al. (2021). Kelemahan metode backpropagation dalam pelatihan jaringan saraf. Jurnal Informatika dan Sistem Cerdas, 4(2), 55–63.

Han, D., et al. (2023). Attention mechanism in neural networks. International Journal of Artificial Intelligence Research, 12(1), 101–115.

Han, L., et al. (2024). Long short-term memory for time-series forecasting. Journal of Machine Learning Applications, 15(2), 88–102.

Jalgaonkar, A., & Kulkarni, P. (2021). Rainfall prediction using support vector machine, random forest, and linear regression. International Journal of Data Science and Analytics, 7(1), 45–58.

Kothai, K., et al. (2023). Computer-based rainfall prediction models. Journal of Environmental Computing, 9(2), 120–133.

Lattifia, L., et al. (2022). Prediksi cuaca menggunakan long short-term memory. Jurnal Teknologi Rekayasa dan Teknologi Informasi, 3(1), 35–42. https://doi.org/10.24843/JTRTI.2022.v03.i01.p35

Le, T., et al. (2020). Rainfall impact and meteorological forecasting. Journal of Climate and Weather Studies, 14(2), 77–89.

Lestari, L., et al. (2023). Keterbatasan backpropagation dalam prediksi. Jurnal Sistem Cerdas Indonesia, 6(1), 30–39.

Nandini, N., et al. (2022). Random forest for rainfall prediction. International Journal of Forecasting Systems, 8(4), 144–156.

Patriardian, F., et al. (2023). Peramalan curah hujan di Provinsi Aceh menggunakan model ARMA. Majalah Ilmiah Matematika dan Statistika, 23(1), 1–14.

Raju, R., et al. (2022). Support vector machine challenges in rainfall prediction. International Journal of Computational Intelligence, 11(2), 60–71.

Rauf, R., et al. (2019). Rainfall, climate factors, and disease prevention. Journal of Environmental Health Studies, 10(1), 22–31.

Sain, S., et al. (2021). Support vector machine for prediction and classification. Journal of Computer Science Applications, 18(3), 95–107.

Secchin, S., et al. (2020). Regression model interpretability. Journal of Statistical Computing, 5(2), 50–62.

Sekhar, P., & Meghana, M. (2020). Backpropagation algorithm for multilayer neural networks. International Journal of Neural Computing, 9(1), 15–28.

Setiawan, S., & Barokah, B. (2022). Prediksi intensitas curah hujan berdasarkan data historis. Jurnal Meteorologi Indonesia, 7(2), 66–74.

Shahid, S., et al. (2021). Rainfall and disease spread prevention. Environmental Health Review, 13(1), 40–51.

Shali, S., et al. (2020). Attention-based LSTM for time-series prediction. Knowledge-Based Systems, 199, Article 105964. https://doi.org/10.1016/j.knosys.2019.05.028

Singh, N., et al. (2019). Weather forecasting using meteorological data. International Journal of Atmospheric Science, 11(4), 210–222.

Suleman, S., & Shridevi, S. (2022). Rainfall prediction for extreme weather mitigation. Journal of Climate Risk Management, 8(2), 55–67.

Sunardi, S., et al. (2020). Prediksi curah hujan menggunakan jaringan saraf tiruan backpropagation. Jurnal Teknologi Informasi dan Ilmu Komputer, 10(2), 155–162. https://doi.org/10.21456/vol10iss2pp155-162

Wahyudi, W., & Pujiastuti, P. (2022). Keterbatasan metode backpropagation pada data pelatihan. Jurnal Teknologi Informasi Modern, 5(1), 25–34.

Xie, X., et al. (2019). Rainfall prediction and ocean health evaluation. Marine Environmental Research, 145, 88–99.

Yusuf, M., Setyanto, A., & Aryasa, K. (2022). Prediksi curah hujan bulanan di Kota Sorong menggunakan multiple linear regression. Jurnal Teknologi Informasi dan Komunikasi, 6(1), 15–24.

Zou, Z., et al. (2022). Limitations of linear regression in nonlinear prediction. Journal of Data Analytics and Modeling, 10(3), 78–89.

Zulkipli. (2023). Produksi padi Jawa Timur tahun 2023. Badan Pusat Statistik Provinsi Jawa Timur.