P-ISSN: 2685-8738 E-ISSN: 2685-872X

Variance: Journal of Statistics and Its Applications
Volume 6 Issue 2 | October 2024
10.30598/variancevol6iss2page175-182

LSTM MODELING WITH AN AUTOREGRESSIVE APPROACH FOR DAILY TEMPERATURE PREDICTION IN GRESIK REGENCY

Azlia Septy Ifadah[1], Brina Miftahurrohmah[2]*, Putri Amelia[3], Ardhi Dwi Firmansyah[4]
[1,2,3,4]Information System Study Program, Universitas Internasional Semen Indonesia
Jl. Veteran, Gresik, 61122, East Java, Indonesia

Corresponding Author's E-mail: *brina.miftahurrohmah@uisi.ac.id*

Abstract: *The zero-hunger program is one of the primary goals of the SDGs, especially in large countries like Indonesia, where hunger remains a serious issue. The agricultural sector plays a crucial role in addressing this problem. However, the effectiveness of this sector is highly dependent on climate changes, such as temperature. Therefore, this research aims to develop a daily temperature prediction model in Gresik Regency using the LSTM method with an autoregressive approach. This model is expected to assist farmers in optimizing planting and harvesting times. The autoregressive approach is applied by analyzing the ACF and PACF plots to determine the lags used as lookback parameters. The research results show that the LSTM model with five lookbacks and 150 epoch parameters provides the best outcomes, with an RMSE value of 0.50, MAE of 0.39, R2 of 0.69, and MAPE of 0.01.*

Keywords: *Autoregressive, Daily Temperature, LSTM, Prediction.*

1. INTRODUCTION

The United Nations (UN) introduced the Sustainable Development Goals (SDGs) in 2015, succeeding the Millennium Development Goals (MDGs) [1]. The SDGs aim to enhance human well-being across various dimensions, encompassing social, economic, and environmental aspects. Despite its global design, SDGs' implementation demands active participation from every stakeholder at both national and regional levels [2]. One primary focus of the SDGs is the zero hunger program, which poses a significant challenge in various nations worldwide, including Indonesia [3].

Gresik Regency, situated in East Java, heavily relies on the agricultural sector as a primary source of income. According to data from the Central Bureau of Statistics of East Java, in 2018, 11,203 residents of Gresik Regency were engaged in farming. However, climate change is one of the sector's most substantial challenges. One aspect of climate change is temperature fluctuation, which has the potential to disrupt agricultural productivity. Its impacts extend beyond yield reduction but also affect food security and farmers' welfare [4], [5], [6].

Amidst these challenges, technology emerges as a pivotal solution, with prominent applications including adopting deep learning, a subfield of machine learning. Deep learning has demonstrated their understanding of complex data patterns, which can help facilitate informed decision-making [7]. One such method within deep learning is Long Short-Term Memory (LSTM), which has proven effective in time series data prediction in previous studies. For instance, Bandung temperature prediction achieved low error rates with the lowest MAPE of 3.84% for maximum temperature and 5.61% for minimum temperature at epoch 10 [8]. Another case is rainfall intensity prediction in Malang Regency, which results in RMSE of 0.98162 and MAE of 0.68847; the consecutive smallest values indicate a high prediction quality [9]. LSTM is a variant of the Recurrent Neural Network (RNN) model that excels in retaining and recalling long-term information, making it commonly used for weather prediction [10]. LSTM model performance analysis evaluates metrics, including RMSE, MAE, R2, and MAPE. Lower RMSE, MAE, and MAPE values indicate superior prediction quality, while an R2 value closer to 1 signifies higher fit accuracy [7]. The autoregressive approach is used as the framework in LSTM model development, wherein autoregressive (AR) assumes that present-period data is influenced by previous-period data [11].

Submitted: 2024-05-23 *Accepted: 2024-10-30*

Hence, this study is conducted to develop a prediction model employing Long-Short-Term Memory (LSTM) with an autoregressive approach. Its goal is to assist farmers in the Gresik Regency in making decisions regarding planting and harvesting, thereby enhancing crop yields, food security, and economic losses for farmers.

2. METHODOLOGY

This research adopts a predictive-quantitative approach using Python and Jupyter tools. Research stages include data collection, pre-processing, exploration and autoregressive modeling, LSTM model development, and model evaluation, as depicted in Figure 1.

Figure 1. Research Stages

2.1. Data Collection

The first step is data collection, accessible through the website https://cds.climate.copernicus.eu/. Data is downloaded in NetCDF format according to the specifications shown in Table 1.

Table 1. Data Download Specifications

Variable	Data Type
Variable	2m temperature
Temperature and pressure	2m temperature
Year	2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023
Month	January, February, March, April, May, June, July, August, September, October, November, December
Day	All Day
Time	All Time
Geographical Area	West 112° East 113° North -7° South -8°

The downloaded data consists of hourly temperature data in Gresik Regency in Kelvin. The total data obtained from the download is 78,888, with the number of data per year shown in Table 2.

Table 2. Annual Total Data

Year	Total
2015	8760
2016	8784
2017	8760
2018	8760
2019	8760
2020	8784
2021	8760
2022	8760
2023	8760

2.2. Data Pre-processing

The first step in the pre-processing stage is to convert the data format from NetCDF per hour to daily CSV for each year. The second step is to convert the data from Kelvin units to Celsius using the formula data (°K) – 278.5. Then, merge all annual data formatted into CSV and adjust the units to Celsius into one CSV file, so the total data now amounts to 3278. Daily temperature data with Celsius units can be seen in Table 3.

Table 3. Daily Temperature Data in Gresik Regency

Date	Temperature
2015-01-01	25.747131
2015-01-02	24.670347
2015-01-03	25.378538
…	…
2023-12-29	27.860222
2023-12-30	27.489792
2023-12-31	27.548430

The next step is normalization to reduce errors using **MinMaxScaler**, where actual data is converted to interval values [0, 1]. The equation used to normalize data is as follows [9].

$$x' = \frac{x - x_{min}}{x_{max} x_{min}} \tag{1}$$

Next, the data is separated into two parts: training data and testing data, with 80% for training data and 20% for testing data. From a total of 3278 data, 2629 data were separated into training and 658 for testing.

2.3. Data Exploration and Autoregressive Modeling

Data exploration is conducted in this stage by analyzing descriptive statistics, including finding the minimum, maximum, mean, median, and standard deviation values to understand the data characteristics. Next, autoregressive modeling is performed by identifying data stationarity; this process includes determining stationarity in mean and variance using time series, autocorrelation function (ACF), and partial autocorrelation function (PACF) plots. A differencing process is applied if the data does not show stationarity. After that, the ACF and PACF plots of stationary data are analyzed to determine the lag, which will be a crucial parameter in the LSTM model. The chosen lag is used as the look-back parameter in the LSTM model. The general form of the autoregressive (AR) model is shown in Equation 2 [12].

$$x_t = \varphi_1 x_{t-1} + k + \varphi_p x_{t-p} + e_t \tag{2}$$

.The look-back parameter determines the number of previous time steps used as input features to predict the next time step. For example, suppose the look-back is set to 3. In that case, each sample in the temperature dataset will use temperature data from the previous three days to predict the current temperature, as shown in Equation 3.

$$x_t = \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \varphi_3 x_{t-3} + e_t \tag{3}$$

2.4. LSTM Model

In developing the LSTM model for processing temperature data, it is essential to determine the appropriate parameters to generate accurate modeling, such as setting the number of hidden layers or appropriate hyperparameters [13]. In this study, researchers employ several schemes with *epoch* parameters, as listed in Table 4.

Table 4. Parameter for LSTM Model

Parameter	Value
Model	*Sequential*
Hidden Layer	4
Dense Layer	1
Optimizer	Adam

Parameter	Value
Loss Function	MSE
Batch Size	1
Epoch	50, 100, 150

The study uses 50, 100, and 150 epoch values to train the LSTM model with daily temperature data. It aims to observe how changes in the number of epochs affect the model's performance in predicting daily temperature data.

2.5. Model Evaluation

The final stage involves model evaluation using RMSE, MAE, R2, and MAPE. The equations used to calculate the evaluation metrics are shown below.

1) *Root Mean Squared Error* (RMSE) [14]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} (\widehat{x_t} - x_t)^2} \tag{4}$$

2) *Mean Absolute Error* (MAE) [9]:

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |\widehat{x_t} - x_t| \tag{5}$$

3) Coefficient of Determination (R2) [15]:

$$R^2 = 1 - \frac{\sum_{t=1}^{n} (x_t - \widehat{x_t})^2}{\sum_{t=1}^{n} \left(x_t - \underline{x}\right)^2} \tag{6}$$

4) Mean Absolute Percentage Error (MAPE) [8]:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n} \left|\frac{x_t - \widehat{x_t}}{x_t}\right| 100\% \tag{7}$$

3. RESULTS AND DISCUSSION

3.1. Data Exploration and Autoregressive Modeling

The daily temperature data for Gresik Regency indicate minimal variability, with a standard deviation significantly lower than the range between the minimum and maximum values, which is 6.2°C, as shown in Table 5.

Table 5. Descriptive Statistic

Criteria	Value
Minimum	23.18
Maximum	29.38
Mean	26.24
Median	26.22
Standard Deviation	0.88

To determine whether the daily temperature data is stationary in both mean and variance, a time series plot and analysis of ACF and PACF plots were utilized. The time series plot used to observe the daily temperature patterns in Gresik Regency is shown in Figure 2.

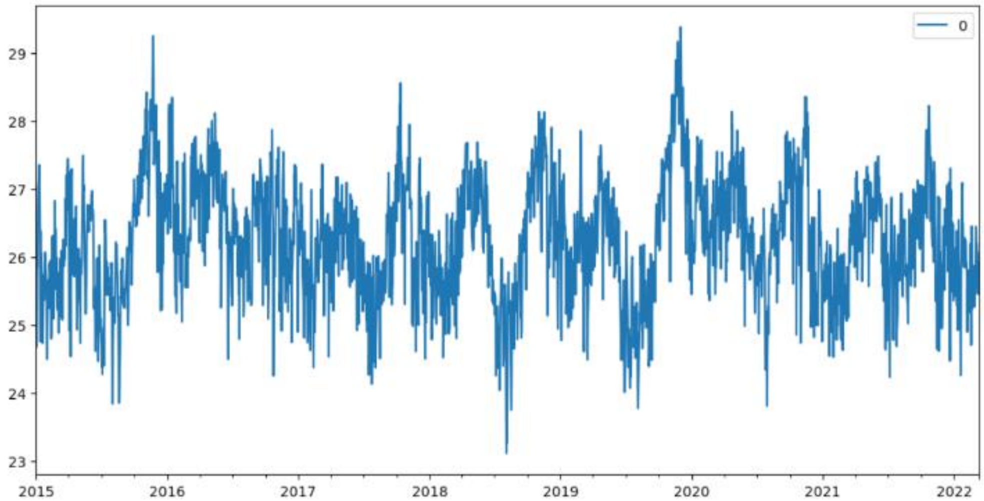

Figure 2. Time Series Plot for Gresik Regency

The time series plot in Figure 2 reveals that the daily temperatures in Gresik Regency generally do not consistently hover around the mean value of 26.24°C, indicating non-stationarity in the mean. Furthermore, the variance of 0.88 is insufficient to confirm stationarity in variance.

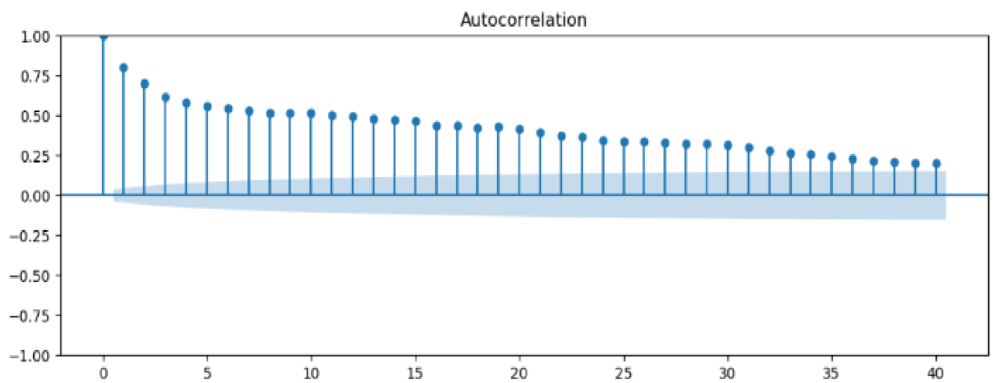

Figure 3. Non-Stationary ACF Plot

Figure 3 shows that the daily temperature data for Gresik Regency is non-stationary, characterized by a slow decay in the ACF plot lags. Therefore, differencing is required to achieve stationarity. The ACF and PACF plots after differencing are shown in Figure 4.

Then, Figure 4 shows that the ACF and PACF plots decline rapidly, with several lags extending beyond the confidence bounds (peaks). The PACF plot indicates that correlations between past and current values are significant up to lag 5, as the lines extending beyond the threshold indicate strong correlations. The correlation results suggest that the values from five previous time points provide crucial information for predicting future values. Beyond lag 5, the correlation diminishes and no longer exceeds the reference threshold, signifying that additional information from subsequent time points is unnecessary. Thus, lags 1 to 5 are selected as the look-back parameters for the LSTM model.

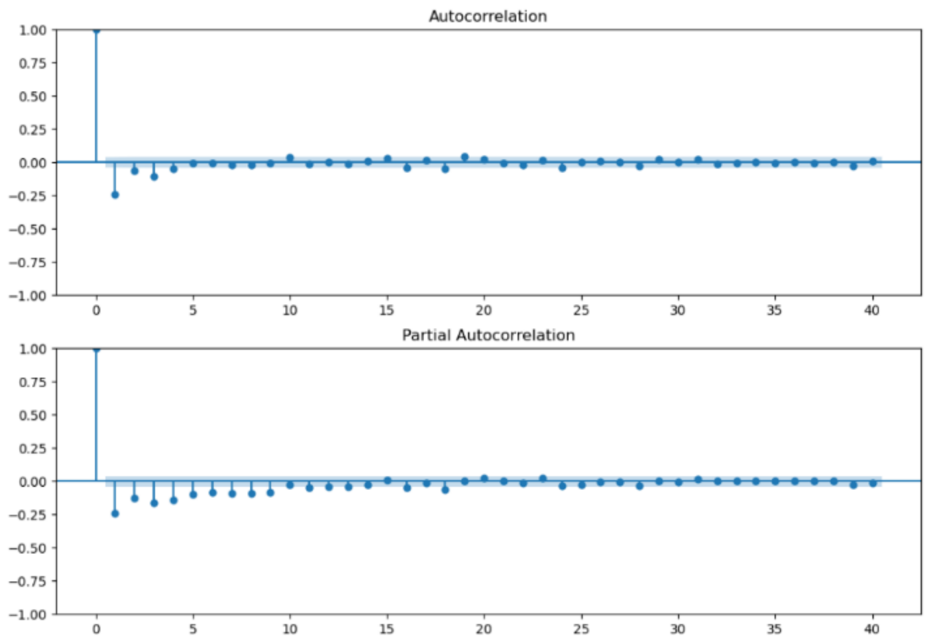

Figure 4. ACF and PACF Plots After Differencing

3.2. LSTM and Model Evaluation

This study employs the LSTM model to predict daily temperatures in the Gresik Regency and initializes the basic parameters used to train the model, which is explained in detail in Table 4. Testing was conducted using combinations of look-back parameters and epochs to evaluate model performance, with the results summarized in Table 6.

Table 6. Scenario Testing Results

Data	*Look Back*	*Epoch*	RMSE	MAE	R2	MAPE
Train	1	50	0.55	0.44	0.60	0.02
Test			0.53	0.43	0.64	0.02
Train		100	0.52	0.41	0.64	0.02
*Test			**0.51**	**0.40**	**0.68**	**0.02**
Train		150	0.52	0.40	0.64	0.02
Test			0.51	0.40	0.67	0.02
Train	2	50	0.53	0.40	0.64	0.02
Test			0.52	0.40	0.66	0.02
Train		100	0.52	0.40	0.64	0.02
*Test			**0.51**	**0.40**	**0.67**	**0.02**
Train		150	0.52	0.40	0.65	0.02
Test			0.51	0.40	0.67	0.02
Train	3	50	0.51	0.40	0.66	0.02
*Test			**0.50**	**0.39**	**0.68**	**0.02**
Train		100	0.52	0.40	0.65	0.02
Test			0.51	0.40	0.68	0.02
Train		150	0.52	0.40	0.66	0.02
Test			0.51	0.40	0.68	0.02
Train	4	50	0.51	0.40	0.66	0.02
*Test			**0.50**	**0.40**	**0.68**	**0.02**
Train		100	0.51	0.40	0.66	0.02
Test			0.51	0.40	0.68	0.02

Data	*Look Back*	*Epoch*	RMSE	MAE	R2	MAPE
Train		150	0.51	0.39	0.66	0.02
Test			0.51	0.39	0.68	0.02
Train		50	0.52	0.41	0.65	0.02
Test			0.51	0.40	0.68	0.02
Train	5	100	0.51	0.40	0.66	0.02
Test			0.50	0.39	0.68	0.02
Train		150	0.51	0.39	0.67	0.01
*Test			**0.50**	**0.39**	**0.69**	**0,01**

(*) indicates the best model for each look-back and epoch combination.

The testing results provide insights into the LSTM model's performance in predicting daily temperatures in the Gresik Regency. Based on the test conducted on the LSTM model with look-back and epochs parameters, it was found that the optimal model for RMSE was achieved by employing a look-back of 3, meaning the model sequentially considers temperatures from one, two, and three days before predicting the current day's temperature and epoch of 50 resulting in an RMSE value of 0.50. For MAE, the best model was also achieved by employing look-back 3 and epoch 50, resulting in an MAE of 0.39. On R2, the best model was obtained with a look-back of 5 and epoch 150, resulting in an R2 of 0.69. This indicates that using the data from one to five days prior could explain 69% of the variability in prediction, with the remaining 31% potentially attributable to other factors not included in the model, such as weather conditions or environmental influences. The best MAPE was also achieved by using a look-back of 5 and epoch 150, with a value of 0.01. This low MAPE indicates that the model has a minimal percentage of error in daily temperature predictions. These findings can serve as a recommendation for future research to consider additional variables to enhance model accuracy further.

4. CONCLUSION

The study results indicate that the daily temperature in Gresik Regency exhibits limited variation, with a low standard deviation, signifying limited diversity. The LSTM model proved effective in predicting daily temperatures, with the best testing results found using a combination of five look-back and *epoch* 150, with evaluation metrics for RMSE of 0.50, MAE of 0.39, R^2 of 0.69, and MAPE of 0.01. Based on these findings, the LSTM model shows potential as an effective tool for future temperature prediction.

REFERENCES

[1] R. S. Shahmohamadloo, C. M. Febria, E. D. G. Fraser, and P. K. Sibley, “The sustainable agriculture imperative: A perspective on the need for an agrosystem approach to meet the United Nations Sustainable Development Goals by 2030,” *Integr. Environ. Assess. Manag.*, no. November 2023, 2021, doi: 10.1002/ieam.4558.

[2] Larizza Dewanty Putri, “PROGRAM SUSTAINABLE DEVELOPMENT GOALS (SDGs) DESA DALAM MENCIPTAKAN KESEJAHTERAAN,” *J. Penelit.*, pp. 1–93, 2023.

[3] C. Saepulmilah *et al.*, “FERMENTASI (PORASI) UNTUK MEMPROMOSIKAN SDG S DALAM MODEL PERTANIAN,” no. c, pp. 27–32, 2023.

[4] N. Herlina and A. Prasetyorini, “Effect of Climate Change on Planting Season and Productivity of Maize (Zea mays L.) in Malang Regency,” *J. Ilmu Pertan. Indones.*, vol. 25, no. 1, pp. 118–128, 2020, doi: 10.18343/jipi.25.1.118.

[5] R. J. Nugroho and A. A. Habiballoh, “Studi Climate Smart Agricultur (CSA) Perubahan Iklim terhadap Ketahanan Pangan,” *J. Pendidik. Tambusai*, vol. 7, pp. 16605–16613, 2023.

[6] E. Surmaini, E. Runtunuwu, and I. Las, “Agriculturals Effort to Anticipate Climate Change,” *J. Penelit.*, vol. 30, no. 98, pp. 1–7, 2017.

[7] A. A. Ningrum, I. Syarif, A. I. Gunawan, E. Satriyanto, and R. Muchtar, “Algoritma Deep Learning-Lstm Untuk Memprediksi Umur Deep Learning-Lstm Algorithm for Power Transformer Lifetime,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 3, pp. 539–548, 2021, doi: 10.25126/jtiik.202184587.

[8] S. Zahroh, Y. Hidayat, R. S. Pontoh, A. Santoso, Sukono, and A. T. Bon, “Modeling and forecasting daily temperature in bandung,” *Proc. Int. Conf. Ind. Eng. Oper. Manag.*, no. November, pp. 406–412, 2019.

[9] T. A. Faisal Muhammad and M. I. Irawan, “Implementasi Long Short-Term Memory (LSTM) untuk Prediksi Intensitas Curah Hujan (Studi Kasus: Kabupaten Malang),” *J. Sains dan Seni ITS*, vol. 12, no. 1, 2023, doi: 10.12962/j23373520.v12i1.106892.

[10] Y. Sudriani, I. Ridwansyah, and H. A Rustini, “Long short term memory (LSTM) recurrent neural network (RNN) for discharge level prediction and forecast in Cimandiri river, Indonesia,” *IOP Conf. Ser. Earth Environ. Sci.*, vol. 299, no. 1, 2019, doi: 10.1088/1755-1315/299/1/012037.

[11] Fitriani, E. T. Herdiani, and M. Saleh, “Pemodelan Autoregressive (AR) pada Data Hilang dan Aplikasinya pada Data Kurs Mata Uang Rupiah,” *J. Mat. Stat. dan Komputasi*, vol. 9, no. 2, pp. 69–85, 2013.

[12] H. Wiguna, Y. Nugraha, F. Rizka R, A. Andika, J. I. Kanggrawan, and A. L. Suherman, “Kebijakan Berbasis Data: Analisis dan Prediksi Penyebaran COVID-19 di Jakarta dengan Metode Autoregressive Integrated Moving Average (ARIMA),” *J. Sist. Cerdas*, vol. 3, no. 2, pp. 74–83, 2020, doi: 10.37396/jsc.v3i2.76.

[13] S. Hidayatullah and A. Cherid, “Prediksi Temperatur Cuaca di Negara Norwegia Menggunakan Metode LSTM,” *Simkom*, vol. 8, no. 2, pp. 187–198, 2023, doi: 10.51717/simkom.v8i2.192.

[14] D. K. A, M. M. A, and S. Schlüter, “Machine Translated by Google Pembelajaran Mesin dengan Aplikasi Prakiraan suhu jangka pendek menggunakan jaringan saraf konvolusional — Aplikasi untuk berbagai stasiun cuaca di Jerman,” vol. 2, no. November, 2020.

[15] M. Diqi and H. Hamzah, “Improving Stock Price Prediction Accuracy with StacBi LSTM,” *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 9, no. 1, pp. 10–26, 2024, doi: 10.14421/jiska.2024.9.1.10-26.