Zero : Jurnal Sains, Matematika, dan Terapan
E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 9, Number 2, 2025
DOI: 10.30829/zero.v9i2.26043
Page: 466-476

Predicting Malaria Incidence Using LSTM and Environmental Variables

[1] **Wellie Sulistijanti**
Institut Teknologi Statistika dan Bisnis Muhammadiyah Semarang
[2] **Laelatul Khikmah**
Institut Teknologi Statistika dan Bisnis Muhammadiyah Semarang
[3] **Erisa Adyati Rahmasari**
Universitas Dian Nuswantoro Semarang
[4] **Cikal Arbitan Putra Sangnandha**
Institut Teknologi Statistika dan Bisnis Muhammadiyah Semarang
[5] **Idan Maulana Yusuf**
Institut Teknologi Statistika dan Bisnis Muhammadiyah Semarang
[6] **Dzahari Alikharimah Azizah**
Institut Teknologi Statistika dan Bisnis Muhammadiyah Semarang

Article Info

Article history:

Accepted, 20 October 2025

Keywords:

Climate change;
Environmental factors;
Long Short-Term Memory;
Malaria incidence forecasting;
Time-series prediction.

ABSTRACT

Climate change is exacerbating malaria risk in Indonesia, especially in Papua. This study proposes a Bidirectional Long Short-Term Memory (LSTM) model to forecast malaria incidence using climate variables. The dataset comprises monthly malaria and climate records (rainfall, temperature, humidity) from four high-endemic provinces between 2014 and 2024. Key methodologies included data augmentation to address data imbalances and a grouped time-series cross-validation for robust model evaluation. An ARIMA model was implemented as a validation baseline to benchmark the proposed approach. The Bi-LSTM model delivered superior performance, achieving an average test R^2 of 0.7210 and SMAPE of 11.02%. the model demonstrated excellent generalization with no evidence of overfitting, significantly outperforming the ARIMA baseline. The findings validate the use of deep learning models as effective tools for public health surveillance, providing reliable early warnings to support timely interventions. Future work will apply SHAP interpretability techniques and expanding the model's geographic scope.

Corresponding Author:

Wellie Sulistijanti,
Study Program of Statistics,
Institut Teknologi Statistika dan Bisnis Muhammadyah Semarang, Indonesia
Email: wellie.sulistijanti@itesa.ac.id

1. INTRODUCTION

Global climate change drives an increase in the population and expansion of the distribution of insect vectors of human diseases, particularly in tropical regions. Specific climatic factors, such as rising temperatures, accelerate the development cycle of the *Plasmodium* parasite within the *Anopheles* mosquito, while fluctuations in rainfall

and humidity directly influence the availability of breeding sites and vector survival rates [1]. As a result, mosquito-borne diseases have emerged as a serious threat to global public health. Current efforts are focused on controlling and ultimately eliminating vector-borne infectious diseases, such as malaria [2]. Malaria remains a significant public health challenge both globally and nationally. In 2023, the World Health Organization (WHO) reported approximately 263 million malaria cases across 83 countries, resulting in an estimated 597,000 deaths worldwide [3]. While most malaria deaths occur in sub-Saharan Africa, the disease remains a persistent threat in other endemic regions, including Southeast Asia [4]. In 2023, Indonesia contributed 27% of Southeast Asia's malaria burden and accounted for more than half of the region's malaria-related deaths [5]

Recent national data reflects a concerning trend. According to the Indonesian Ministry of Health reported a 72% increase in confirmed malaria cases, rising from 304,607 in 2021 to 543,965 in 2024, along with a nearly threefold increase in annual fatalities (from 48 to 132) [6]. Several eastern provinces, including Papua, East Nusa Tenggara, and West Kalimantan, continue to report high endemicity. Among these, Papua holds a unique position, contributing more than 90% of Indonesia's malaria cases despite representing only 2% of the national population. Unlike most regions in Indonesia, where malaria peaks seasonally during the rainy season (December–June), Papua experiences year-round transmission due to continuous rainfall. Recognizing these challenges, The WHO Global Technical Strategy for Malaria 2016–2030 and the Indonesia National Action Plan for Acceleration of Malaria Elimination 2020-2026 highlight the urgent need to strengthen malaria surveillance as a core strategy for accelerating elimination In line with this, computational modelling has emerged as a promising tool to enhance malaria surveillance through better prediction accuracy, early warning capabilities, and targeted response strategies [7]

In recent years, various studies have utilized computational approaches to support malaria prediction. Menda et al. (2021) proposed a hybrid model that combined Machine Learning (ML) techniques with Autoregressive Integrated Moving Average (ARIMA) models to forecast forecasting malaria cases [8]. Similarly, Javaid et al. [9] employed an integrative approach by combining Web-based Geographic Information Systems (WebGIS) with various ML algorithms, including Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), and Multilayer Perceptron (MLP), to conduct spatial-temporal analysis of malaria distribution in Pakistan.

More recently, Naroum et al., (2025) demonstrated that Long Short-Term Memory (LSTM) networks outperformed other ML models, such as SVM, RF, Ridge, Lasso, and ElasticNet, in predicting malaria cases based on climate variables like rainfall and temperature in Cameroon [10]. Their results showed that LSTM achieved the lowest Root Mean Squared Error (RMSE) among all tested models, aligning with findings from [11]who also reported superior accuracy of LSTM over traditional models such as ARIMA and RF. While LSTM has shown strong performance in modeling time-series data for climate-sensitive diseases, its application in the Indonesian context, especially in high-transmission regions like Papua, remains limited. This reveals a significant gap in the current surveillance system, which still relies heavily on retrospective data analysis. In contrast, computational models like LSTM offer a paradigm shift towards proactive forecasting, enabling the generation of timely early warnings. This capability is particularly critical in Papua, which bears over 90% of Indonesia's malaria burden. The region's high transmission rates and logistical challenges mean that data-driven, predictive interventions are essential for pre-positioning resources and preventing localized cases from escalating into large-scale outbreaks.

However, few studies have applied LSTM models to malaria forecasting in Indonesia's high-transmission regions. Therefore, this study aims to address this gap by proposing an integrated LSTM-based model for predicting malaria cases in Indonesia using climate and environmental data. Both malaria incidence data and climate-related variables—including rainfall, humidity, and temperature—were entirely obtained from the Statistics Indonesia (Badan Pusat Statistik, BPS) publications at the provincial level. The study focuses on four endemic provinces: Papua, West Papua, East Kalimantan, and East Nusa Tenggara, which have consistently reported high malaria incidence between 2014 and 2024. Climate change influences the transmission of malaria, primarily through temperature and rainfall factors. Dry conditions reduce the population of mosquito vectors, while high rainfall increases water pooling in rice fields and springs, thereby raising the density of Anopheles sp. and contributing to fluctuations in malaria incidence [12]

2. RESEARCH METHOD

We propose a Long Short-Term Memory (LSTM) model to predict malaria transmission across endemic regions in Indonesia by integrating climatic and environmental variables. Long Short-Term Memory (LSTM) is specifically designed to effectively handle and process time-series data, which is particularly relevant for predicting phenomena with temporal dependencies, such as malaria transmission influenced by seasonal and climatic factors [13]. The illustration of the research flow can be seen in Figure 1.

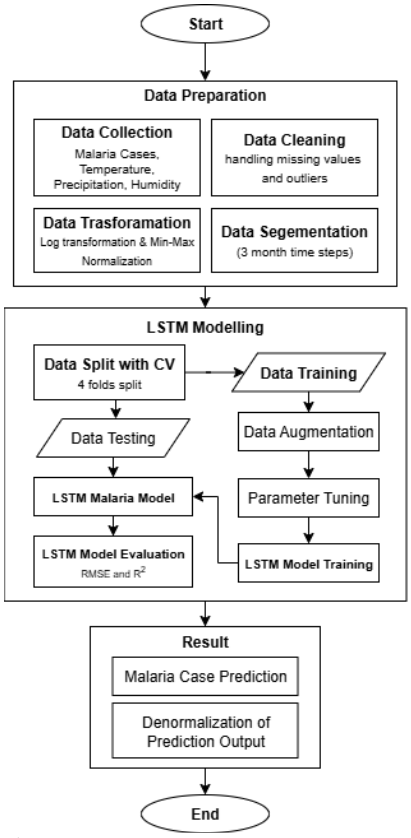

Figure 1. Research Methodology

2.1 Data Preparation

The dataset for this study consists of monthly time-series data from 2014–2024, entirely obtained from Statistics Indonesia (Badan Pusat Statistik, BPS). It includes monthly malaria case records along with corresponding climate variables—rainfall, temperature, and humidity—compiled at the provincial level for model development and analysis. The full list of indicators and data sources used in this study is summarized in Table 1.

Tabel 1. Indicator and Malaria Data Transmission

No	Indicator	Variabel	References
1	Malaria Transmission	Number of malaria cases per region	[14]
		Historical monthly malaria case data	[15], [16], [17], [18]
2	Environment Factor	Average temperature (°C)	[19], [20], [21], [22]
		Precipitation (mm)	[19], [20], [22], [23]
		Relative humidity (%)	[19], [20], [21]

In order to ensure the suitability of the dataset for model training and prediction, a comprehensive data preprocessing stage was conducted. The intial step involved a thorough examination of the dataset for completeness and the presence of outliers. Following the integrity check, outliers were addressed using the Interquartile Range (IQR) capping method. Since certain variables such as malaria cases and rainfall exhibited right-skewed distributions, a logarithmic transformation was first applied to reduce skewness and lessen the impact of extreme outliers.

Given that input features such as temperature, rainfall, and humidity vary in scale, Min-Max normalization was applied to rescale all variables to a common range between 0 and 1 [10] as shown in Eq. (4):

$$x' = \frac{x - \min x}{\max(X) - \min(X)} \tag{4}$$

where x refers to the original data value, X the set of all values used for normalization, and x' represents the resulting normalized value.

To address class imbalance across provinces, data augmentation techniques were introduced prior to sequence segmentation. Data augmentation aims to increase the amount and variety of this data, thereby allowing the model to better generalize the data and recognize features in the data that it has never seen before, as well as

preventing overfitting. Two approaches were employed: (i) oversampling of minority provinces using interpolation-based synthetic generation, and (ii) conservative augmentation, which included the injection of small gaussian noise and magnitude warping [24]. Gaussian noise operates by adding random values drawn from a normal distribution to each point in the signal, expressed as Eq. (5):

$$y_t = x_t + \epsilon,\ \epsilon \sim N(0, \sigma^2) \tag{5}$$

Meanwhile, magnitude warping modifies the temporal series by applying smoothly varying scaling factors, formulated in Eq. (6):

$$x'_t = x_t \cdot \alpha_t,\ \alpha_t = N(1, \sigma^2) \tag{6}$$

where x'_t denotes the original signal at time t, ϵ represents Gaussian noise, and α_t is a smooth warping curve interpolated from Gaussian-distributed nodes [25]Gaussian noise thereby perturbs individual points in a controlled manner, while magnitude warping randomly scales certain segments of the series, both aiming to enrich the training set while preserving statistical properties and temporal dynamics.

Following that, the time-series data were segmented using a sliding window with a sequence length of 3 months. This duration was selected to reflect the seasonal nature of malaria transmission, which is closely linked to variations in rainfall, temperature, and humidity. A 3-month window allows the model to capture relevant short-term trends and identify recurring seasonal patterns. Reserch from [26]demonstrated that malaria incidence often follows quarterly climatic cycles, with spikes typically occurring during or shortly after the rainy season. Thus, this approach enhances the model's ability to anticipate periods of increased transmission.

2.2 Deep Learning Modelling using LSTM

The LSTM unit comprises three gates that regulate the flow of information through the memory cell: the forget gate, input gate, and output gate [13]These gates are mathematically formulated as follows:

a. Forget Gate: Determines which information from the previous cell state should be discarded using Eq. (7).

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{7}$$

where W_f , b_f is weight matrix and bias of the forget gate, h_{t-1} is previous hidden state at time t-1, x_t is input at time t and σ is sigmoid activation function.

b. Input Gate: Updates the cell state with new candidate information using sigmoid function (Eq. 8) and tanh activation (Eq. 9).

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{8}$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{9}$$

where W_i , b_i is weight matrix and bias of the input gate and W_c , b_c is weight matrix and bias of the candidate cell state

c. Output Gate: Determines the final output by applying a sigmoid function (Eq. 10) which is then multiplied with the tanh-transformed cell state to generate the new hidden state (Eq. 11).

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{10}$$
$$h_t = o_t \times \tanh(C_t) \tag{11}$$

with W_o, b_o is weight and bias of the output gate.

The architecture of the LSTM model used in this study follows a sequential structure. It consists of a single bidirectional LSTM layer with 12 hidden units, followed by a dropout layer (rate 0.4), a L2 regularization layer, and a dense output layer with linear activation. This configuration was chosen to balance complexity and generalizability in modeling disease-related time-series data [13], [14]. The model was trained using the Adam optimizer with a learning rate of 0.0005, over 150 epochs with a batch size of 8. Additionally, an earlystopping callback was used in this training to control the training. Such architecture ensures the model is robust enough to capture seasonal fluctuations in malaria cases, while remaining replicable for future studies (Naroum et al., 2025). After constructing the structure of LSTM model as illustrated in Figure 1, the pre-processed dataset is chronologically split using grouped time series cross validation method with 4 folds. This cross-validation approach allows the model to be evaluated on unseen data and helps mitigate overlearning [10]. The training phase produces initial prediction of malaria cases that can support decision-making process. These findings facilitate the

identification of temporal distributions that influence malaria transmission. Furthermore, to validate its effectiveness, the performance of the proposed LSTM model is benchmarked against ARIMA (Autoregressive Integrated Moving Average), a classical statistical model widely used for time-series forecasting.

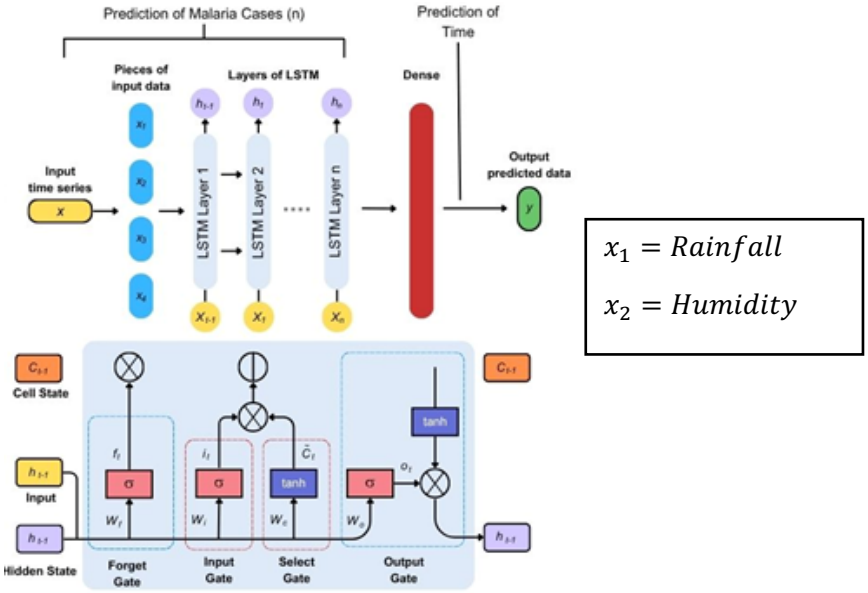

Figure 2. LSTM Model for Malaria Prediction

Rainfall, humidity, and temperature are the main factors in this malaria prediction model because all three directly affect the life of the Anopheles mosquito [27]. Rainfall creates puddles, which are breeding grounds for mosquitoes. High humidity helps mosquitoes survive longer, increasing the chances of transmission [28]. Meanwhile, the optimal temperature accelerates the development of malaria parasites in the mosquito's body. The LSTM model uses these data to analyze the complex relationship between environmental conditions and the number of malaria cases.

2.3 Baseline Model using ARIMA

To validate the performance of the proposed Bi-LSTM model, an ARIMA (Autoregressive Integrated Moving Average) model was implemented as a classical statistical baseline. ARIMA is a widely-used statistical method for time-series forecasting that models a variable's future values based on its own past values, specifically its lags (Autoregressive) and lagged forecast errors (Moving Average) [1]. It was selected to serve as a strong linear, sequential benchmark, providing a contrast to the non-linear capabilities of the LSTM. Unlike the multivariate LSTM, a univariate ARIMA model was fitted individually to the time series of malaria cases for each province.

2.4 Experimental Setup

All models were implemented in the Python programming language (version 3.12), utilizing the TensorFlow (version 2.19) and Statsmodels libraries. The key hyperparameters for the Bi-LSTM model were selected based on common practices for time-series forecasting and preliminary experiments to balance performance and model complexity. The final architecture consisted of a Bi-LSTM layer with 12 units, an embedding dimension of 4 for provincial data, a dropout rate of 0.4, and L2 regularization of 0.005. The model was trained using an Adam optimizer with a learning rate of 0.0005.

2.5 Evaluation Metrics

The model's predictive performance was evaluated using the test set, based on three standard metrics: Root Mean Squared Error (RMSE), Symmetric Mean Absolute Percentage Error (SMAPE), and the coefficient of determination (R^2) [29]. Here, n represents the total number of malaria cases, y_i denotes the actual value of malaria, and $\hat{y}_i$ denotes the predicted malaria value.

RMSE

RMSE reflects the average magnitude of prediction errors by computing the square root of the mean squared differences between predicted and actual values. This metric assigns greater weight to larger prediction errors, making it particularly effective in detecting substantial deviations between predicted and actual values. The RMSE is computed using the formulation presented in Eq. (12).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i-1}^{n}\left(y_i - \hat{y}_i\right)^2} \quad (12)$$

sMAPE

sMAPE is utilized to measure the predictive accuracy of the forecasting model. As a modification of MAPE, sMAPE provides a symmetric evaluation by normalizing the absolute difference against the mean of the predicted and actual values. This characteristic makes it particularly robust for datasets containing zero or near-zero values, such as the malaria incidence data in this study, thus preventing the generation of misleadingly large percentage errors. The sMAPE value is computed using the formulation presented in Eq. (13).

$$sMAPE = \frac{100\%}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{(|A_t| + |F_t|)/2} \quad (13)$$

R^2

R^2 is employed to assess how closely the model's predictions approximate the actual value [30] The R^2 value is computed using the formulation provided in Eq. (14).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (14)$$

Predicted malaria case

The LSTM model was used to predict malaria cases. After prediction, denormalization was performed to convert the results back to their original scale. The output from the LSTM model will be used to generate malaria risk distribution maps and identify high-risk endemic areas in Indonesia

3. RESULT AND ANALYSIS

Before evaluating the model performance, we provide an overview of the input variables across the four endemic provinces (2014–2024). Table 2 summarizes the mean, minimum, maximum, and standard deviation for each variable.

Table 2. Data of Malaria Cases

Variable	Mean	Min	Max	Std.Dev
Malaria Cases	70001.35	0	78528	12248.86
Rainfall (mm)	201.27	0	7929	375.39
Temperature (°C)	27.04	18.06	38.30	1.79
Humidity (%)	81.53	57.00	97.19	5.29

Based on table 2 presented, malaria cases have a very high daily average, namely 70,001.35 cases, with a maximum value of 78,528, indicating a significant spike in cases. This variation is supported by a large standard deviation, which is 12,248.86, indicating drastic fluctuations. Rainfall, with an average of 201.27 mm and a maximum value of up to 7,929 mm, also showed similar variations, characterized by a high standard deviation of 375,39. On the other hand, temperature and humidity showed relatively more stable values, with an average temperature of 27.04°C (std. dev. 1.79) and an average humidity of 81.53% (std. dev. 5.29), indicating environmental conditions that were consistent enough to support the life of malaria-spreading mosquitoes.

Figure 3 show the comparison of malaria cases from 2014 to 2024 in four provinces (East Kalimantan, NTT, Papua, and West Papua), significant difference in the number of malaria cases between provinces. Papua province consistently recorded the highest number of malaria cases throughout the period, with the number of cases far exceeding the other three provinces, as seen from the green line at the very top level. Meanwhile, East Kalimantan showed the lowest number of cases, with a stable blue line at the very bottom of the graph. West Papua and East Nusa Tenggara (NTT) are in the middle, with West Papua generally having a higher number of cases than NTT, although there have been periods where NTT's cases have increased sharply (for example, in early 2017) or West Papua has seen a drastic decline (around 2022)

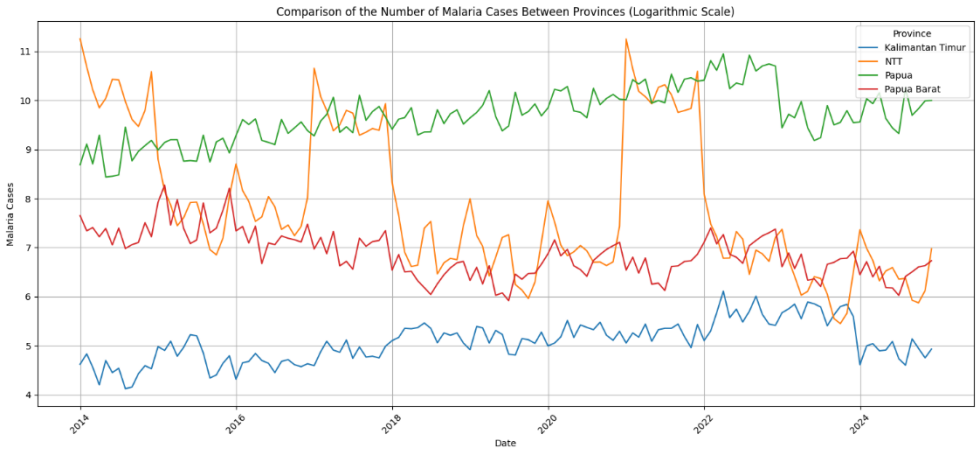

Figure 3. Plot of the monthly incidence of Malaria Cases in Indonesia from 2014 to 2024

Based on graphs related to rainfall data and malaria cases in the four provinces from 2014 to 2024, it can be seen that there is a relationship between the two variables. In East Nusa Tenggara Province, the graph shows a strong correlation. Papua and West Papua are quite significant. Meanwhile, in East Kalimantan, the increase in rainfall in several years was also followed by an increase in malaria cases.

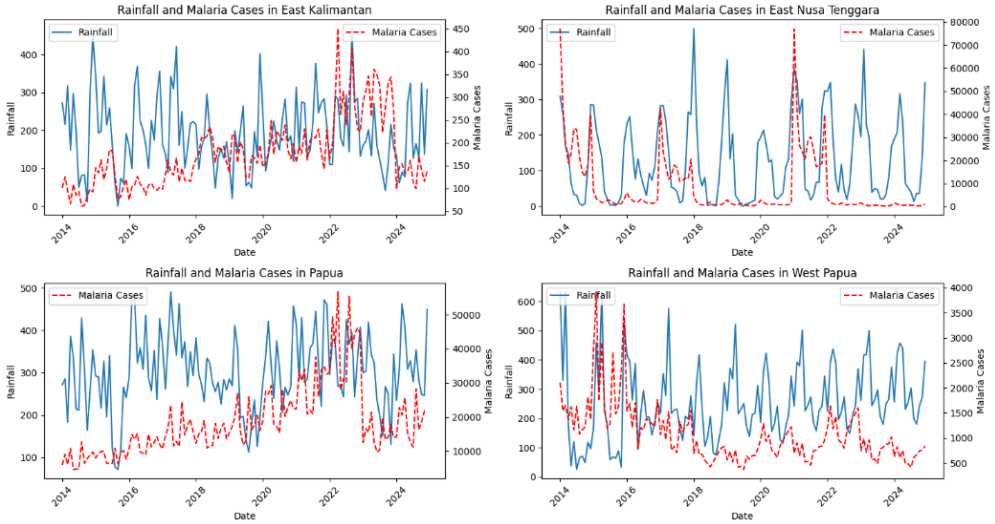

Figure 4. Plot of the Rainfall and Malaria Case Data in each Provinces from 2014 to 2024

3.1 ARIMA Benchmark Results

To validate the superiority of the proposed Bi-LSTM model, we implemented ARIMA as a baseline statistical approach for time-series forecasting. Table 3 presents the performance of ARIMA across the four endemic provinces, evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Symmetric Mean Absolute Percentage Error (sMAPE).

Table 3. ARIMA Performance across Provinces

Province	ARIMA Model	MSE	RMSE	sMAPE
Papua	ARIMA (1,1,1)	471.2020	21.71	25.21%
West Papua	ARIMA (2,0,0)	0.5491	0.74	45.66%
East Nusa Tenggara	ARIMA (1,2,0)	1.5431	1.24	72.34%
East Kalimantan	ARIMA (1,1,1)	11.5050	3.39	75.19%

The performance of the ARIMA model demonstrates a fundamental limitation of this approach to estimating malaria incidence, namely its univariate nature. As a univariate model, ARIMA forecasts future malaria cases based solely on their own past values (lags and historical errors), making it inherently blind to the influence of external, exogenous variables like climatic factors.

East Nusa Tenggara (NTT) has high data variability, with the ARIMA model producing a relatively small MSE (1.5431) but a high sMAPE (72.34%). This indicates that although ARIMA captures some patterns, it

struggles to handle large fluctuations in the data, resulting in relatively high prediction errors. These peaks and troughs are likely driven by climatic events, which the ARIMA model cannot account for. Conversely, in Papua, the model produces a large MSE (471.2020) due to the high volume of cases, but its lower sMAPE (25.21%) suggests it can capture the general baseline incidence. However, it still struggles to model the variability around this mean, as this variability is also heavily influenced by climate.

Ultimately, ARIMA's inability to incorporate multivariate inputs like rainfall and temperature means it cannot capture the complex, non-linear relationships that drive malaria transmission. This underscores the necessity for more advanced models capable of integrating these crucial external drivers.

3.2 Bi-LSTM Model Performance

The evaluation results demonstrated consistent performance across all validation splits. Across the validation splits, the model's average performance on unseen data yielded a test R^2 of 0.7210±0.110, which was superior to its performance on the training data (train R^2 of 0.6702±0.051). This pattern was reinforced by the error metrics, where the average test RMSE of 1.0439 was lower than the train RMSE of 1.2384. Superior performance on the test set strongly indicates that the model successfully learned generalizable patterns rather than memorizing the training data. These results suggest that the LSTM model successfully captured the underlying temporal dynamics of malaria cases while maintaining robustness across different folds.

Table 4. The Performance of LSTM Model

Fold	Test R^2	Train R^2	Test RMSE	Train RMSE	Test sMAPE	Train sMAPE
Split 1	0.7036	0.5902	1.0525	1.4417	10.68%	17.20%
Split 2	0.5516	0.6727	1.4301	1.2110	15.20%	13.26%
Split 3	0.7867	0.6856	0.9595	1.1949	9.35%	12.92%
Split 4	0.8421	0.7322	0.7335	1.1059	8.85%	11.55%
Average	0.7210	0.6702	1.0439	1.2384	11.02%	13.73%

Further evidence of the model's robust generalization is provided by the training and validation loss curves, as depicted in Figure 5. Across the training epochs, both the training loss and the validation loss consistently decreased and converged towards similarly low values. Crucially, the validation loss did not exhibit a significant upward trend, which would typically indicate overfitting. This convergence pattern visually confirms that the regularization techniques employed (dropout and L2 regularization) were effective in preventing the model from merely memorizing the training data, further supporting its ability to learn generalizable temporal patterns.

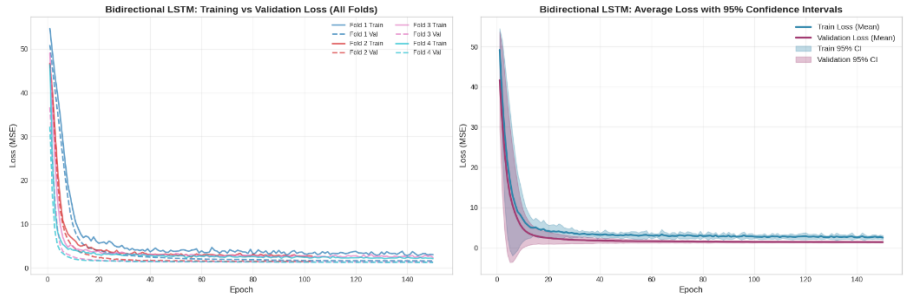

Figure 5. Comparison of Actual Data vs. LSTM Predictions of Malaria Cases in the Last Fold

Following the individual analysis of each model, a direct comparative evaluation confirms the superiority of the Bidirectional LSTM architecture for this forecasting task, with aggregated results presented in Table 4. As established above, the ARIMA model demonstrated limitations in consistently handling the high variability and non-linear patterns inherent in the malaria data. In contrast, the Bi-LSTM model achieved significantly better overall performance, yielding a lower prediction error, with a Test RMSE of 1.0439 (vs. 6.77) and Test sMAPE of 11.02%. Furthermore, a paired t-test comparing the models' prediction errors confirmed this performance gap is statistically significant ($p = 0.0023$). This performance gap underscores the Bi-LSTM's advanced ability to capture the complex, non-linear temporal dependencies between climatic factors and malaria incidence, a critical feature that traditional linear models like ARIMA cannot fully address.

To further explore the performance of the LSTM model, a visual comparison of its prediction results with the actual data from the final fold is presented in Figure 6.

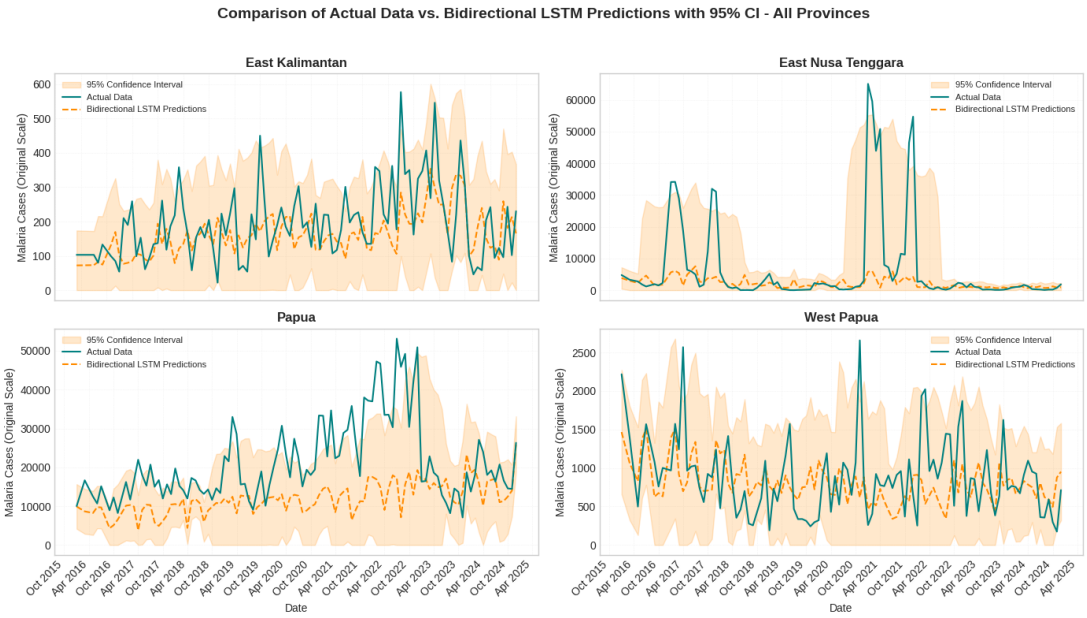

Figure 6. Comparison of Actual Data vs. LSTM Predictions of Malaria Cases in the Last Fold

Furthermore, visual comparisons between actual and predicted values in the final fold provided additional insights into model performance. Four representative plots for East Kalimantan, East Nusa Tenggara, Papua, and West Papua from 2023 to 2024 revealed that the model effectively reproduced the overall trends of malaria cases. Although certain deviations were observed, the predicted series closely followed the actual data patterns, highlighting the model's robustness in capturing temporal and regional variations. Unlike prior Indonesian studies that primarily relied on conventional regression or static spatial analysis, this approach integrates time-series learning with regional differentiation, offering a more adaptive and scalable framework for forecasting malaria incidence across diverse provinces.

4. CONCLUSION

This study applied Long Short-Term Memory (LSTM) networks to forecast malaria cases in four Indonesian provinces using climatic and environmental variables. The results revealed substantial variability in malaria incidence, with Papua consistently recording the highest number of cases, while East Kalimantan showed the lowest. Through comparative analysis, it was found The Bi-LSTM reduced prediction error by 60% compared to ARIMA, demonstrating the feasibility of deep learning for malaria surveillance. The proposed model achieved strong average performance across all cross-validations with an average Test R^2 of 0.7210 and a Test sMAPE of 11.02%. The model results demonstrate excellent generalization ability with no evidence of overfitting.

While previous Indonesian studies predominantly employed conventional ARIMA or standard LSTM models for malaria forecasting, this study introduces a more advanced Bidirectional LSTM framework that integrates data augmentation and grouped time-series cross-validation. These additions enhance the model's ability to handle data imbalance and capture complex, non-linear temporal dependencies between climate variables and malaria incidence. By combining forward and backward temporal learning, the proposed Bi-LSTM effectively models lagged climatic effects that traditional approaches fail to represent, marking a methodological improvement in malaria prediction within Indonesia's climatic context.

The model effectively captures the time lag between climatic events, such as rainfall, and the subsequent rise in malaria cases several weeks later, enabling actionable early warnings in high-risk regions like Papua. This LSTM-based forecasting framework serves as a valuable decision-support tool for public health authorities. Integrating it into routine malaria surveillance could generate early alerts weeks in advance, allowing timely vector control, optimized resource allocation, and strengthened community-level prevention. Ultimately, this approach supports Indonesia's malaria elimination goals through data-driven and proactive interventions.

ACKNOWLEDGEMENTS

This research is fully supported by the Research Grant of the Directorate of Research and Community Service, Ministry of Higher Education, Science, and Technology of the Republic of Indonesia under Contract Numbers 077/LL6/PL/AL.04/2025 and A.1/11/062010/LPPM/V/2025.

5. REFERENCES

[1] U. M. Sirisha, M. C. Belavagi, and G. Attigeri, "Profit Prediction Using ARIMA, SARIMA and LSTM Models in Time Series Forecasting: A Comparison," *IEEE Access*, vol. 10, pp. 124715–124727, 2022, doi: 10.1109/ACCESS.2022.3224938.

[2] T. Santosh, D. Ramesh, and D. Reddy, "LSTM based prediction of malaria abundances using big data," *Comput Biol Med*, vol. 124, p. 103859, Sep. 2020, doi: 10.1016/j.compbiomed.2020.103859.

[3] World Health Organization, *World Malaria Report 2021*. 2021.

[4] S.-X. Zhang *et al.*, "Global, regional, and national burden of malaria, 1990–2021: Findings from the global burden of disease study 2021," *Decoding Infection and Transmission*, vol. 2, p. 100030, 2024, doi: 10.1016/j.dcit.2024.100030.

[5] M. Delvina, E. Yuniarti, E. Barlian, and L. Handayuni, "Trends in Malaria cases by Plasmodium type in West Sumatra province: Analysis of 2017-2023," *Multidisciplinary Science Journal*, vol. 7, no. 10, p. 2025492, Apr. 2025, doi: 10.31893/multiscience.2025492.

[6] D. N. Aisyah *et al.*, "The Changing Incidence of Malaria in Indonesia: A 9-Year Analysis of Surveillance Data," *Adv Public Health*, vol. 2024, no. 1, Jan. 2024, doi: 10.1155/adph/2703477.

[7] M. F. A. K. E. R. R. P. W. P. S. S. Wiwik Anggraeni, "Combination of BERT and Hybrid CNN-LSTM Models for Indonesia Dengue Tweets Classification," *International Journal of Intelligent Engineering and Systems*, vol. 17, no. 1, pp. 813–826, Feb. 2024, doi: 10.22266/ijies2024.0229.68.

[8] D. M. Menda *et al.*, "Forecasting Confirmed Malaria Cases in Northwestern Province of Zambia: A Time Series Analysis Using 2014–2020 Routine Data," *Adv Public Health*, vol. 2021, pp. 1–8, Oct. 2021, doi: 10.1155/2021/6522352.

[9] M. Javaid, M. Sarfraz, M. Aftab, Q. Zaman, H. Rauf, and K. Alnowibet, "WebGIS-Based Real-Time Surveillance and Response System for Vector-Borne Infectious Diseases," *Int J Environ Res Public Health*, vol. 20, no. 4, p. 3740, Feb. 2023, doi: 10.3390/ijerph20043740.

[10] E. Naroum *et al.*, "Comparative analysis of deep learning and machine learning techniques for forecasting new malaria cases in Cameroon's Adamaoua region," *Intell Based Med*, vol. 11, p. 100220, 2025, doi: 10.1016/j.ibmed.2025.100220.

[11] H. Alizadegan, B. Rashidi Malki, A. Radmehr, H. Karimi, and M. A. Ilani, "Comparative study of long short-term memory (LSTM), bidirectional LSTM, and traditional machine learning approaches for energy consumption prediction," *Energy Exploration and Exploitation*, vol. 43, no. 1, pp. 281–301, Jan. 2025, doi: 10.1177/01445987241269496.

[12] K. Tanah Bumbu Kalimantan Selatan *et al.*, "Pengaruh curah hujan, kelembaban, dan temperatur terhadap prevalensi Malaria di The effect of rainfall, humidity, and temperature on malaria prevalence in Tanah Bumbu District South Kalimantan," *JHECDs*, vol. 3, no. 1, pp. 22–27, 2017, doi: 10.22435/jhecds.v3i1.5063.22-27.

[13] S. Raheja and S. Malik, "Prediction of Air Quality Using LSTM Recurrent Neural Network," *International Journal of Software Innovation*, vol. 10, no. 1, pp. 1–16, Apr. 2022, doi: 10.4018/IJSI.297982.

[14] Y. S. Kitawa and Z. G. Asfaw, "Space-time modelling of monthly malaria incidence for seasonal associated drivers and early epidemic detection in Southern Ethiopia," *Malar J*, vol. 22, no. 1, p. 301, Oct. 2023, doi: 10.1186/s12936-023-04742-9.

[15] W. Haileselassie *et al.*, "International border malaria transmission in the Ethiopian district of Lare, Gambella region: implications for malaria spread into South Sudan," *Malar J*, vol. 22, no. 1, pp. 1–10, 2023, doi: 10.1186/s12936-023-04479-5.

[16] G. J. Gbaguidi, N. Topanou, W. L. Filho, and G. K. Ketoh, "Potential impact of climate change on the transmission of malaria in Northern Benin, West Africa," *Theor Appl Climatol*, vol. 155, no. 5, pp. 3525–3539, 2024, doi: 10.1007/s00704-023-04818-1.

[17] M. Javaid, M. S. Sarfraz, M. U. Aftab, Q. uz Zaman, H. T. Rauf, and K. A. Alnowibet, "WebGIS-Based Real-Time Surveillance and Response System for Vector-Borne Infectious Diseases," *Int J Environ Res Public Health*, vol. 20, no. 4, 2023, doi: 10.3390/ijerph20043740.

[18] L. Demoze, F. Gubena, E. Akalewold, H. Brhan, T. Kifle, and G. Yitageasu, "Spatial, temporal, and spatiotemporal cluster detection of malaria incidence in Southwest Ethiopia," *Front Public Health*, vol. 12, no. January, 2024, doi: 10.3389/fpubh.2024.1466610.

[19] W. Haileselassie *et al.*, "International border malaria transmission in the Ethiopian district of Lare, Gambella region: implications for malaria spread into South Sudan," *Malar J*, vol. 22, no. 1, pp. 1–10, 2023, doi: 10.1186/s12936-023-04479-5.

[20] G. J. Gbaguidi, N. Topanou, W. L. Filho, and G. K. Ketoh, "Potential impact of climate change on the transmission of malaria in Northern Benin, West Africa," *Theor Appl Climatol*, vol. 155, no. 5, pp. 3525–3539, 2024, doi: 10.1007/s00704-023-04818-1.

[21] N. Mahdizadeh Gharakhanlou, M. S. Mesgari, and N. Hooshangi, "Developing an agent-based model for simulating the dynamic spread of Plasmodium vivax malaria: A case study of Sarbaz, Iran," *Ecol Inform*, vol. 54, no. September, p. 101006, 2019, doi: 10.1016/j.ecoinf.2019.101006.

[22] Y. S. Kitawa and Z. G. Asfaw, "Space-time modelling of monthly malaria incidence for seasonal associated drivers and early epidemic detection in Southern Ethiopia," *Malar J*, vol. 22, no. 1, pp. 1–13, 2023, doi: 10.1186/s12936-023-04742-9.

[23] D. Diriba, S. Karuppannan, T. Regasa, and M. Kasahun, "Spatial analysis and mapping of malaria risk areas using geospatial technology in the case of Nekemte City, western Ethiopia," *Int J Health Geogr*, vol. 23, no. 1, 2024, doi: 10.1186/s12942-024-00386-3.

[24] L. Roque, C. Soares, V. Cerqueira, and L. Torgo, "L-GTA: Latent Generative Modeling for Time Series Augmentation," 2025. doi: XXXXXXX.XXXXXXX.

[25] T. Žvirblis, A. Pikšrys, D. Bzinkowski, M. Rucki, A. Kilikevičius, and O. Kurasova, "Data Augmentation for Classification of Multi-Domain Tension Signals," *Informatica*, pp. 883–908, Nov. 2024, doi: 10.15388/24-INFOR578.

[26] I. E. Rozi *et al.*, "Rapid entomological assessment in eight high malaria endemic regencies in Papua Province revealed the presence of indoor and outdoor malaria transmissions," *Sci Rep*, vol. 14, no. 1, p. 14603, Jun. 2024, doi: 10.1038/s41598-024-64958-w.

[27] O. Nkiruka, R. Prasad, and O. Clement, "Prediction of malaria incidence using climate variability and machine learning," *Inform Med Unlocked*, vol. 22, p. 100508, 2021, doi: 10.1016/j.imu.2020.100508.

[28] R. Sriviswa, H. Radhika Reddy, N. M. Rajendran, V. Sowmya, and E. A. Gopalakrishnan, "Deep Attention Model for Malaria Prediction," Springer, 2025, pp. 45–56. doi: 10.1007/978-3-031-87154-2_5.

[29] W. SULISTIJANTI and A. C. Vayuanita, "PERAMALAN HASIL PRODUKSI PADI DI PROVINSI JAWA TENGAH MENGGUNAKAN METODE HYBRID SARIMA-FUZZY TIME SERIES CHEN," *Agritech : Jurnal Fakultas Pertanian Universitas Muhammadiyah Purwokerto*, vol. 25, no. 2, p. 194, Jul. 2024, doi: 10.30595/agritech.v25i2.21835.

[30] R. Syabrina and G. Wang, "Predictive Analysis of Malaria Cases in Indonesia Using Machine Learning," *Syntax Literate ; Jurnal Ilmiah Indonesia*, vol. 9, no. 9, pp. 4849–4859, Sep. 2024, doi: 10.36418/syntax-literate.v9i9.16714.