Journal of Applied Data Sciences
Vol. 7, No. 3, September 2026, pp. 2346-2360

A Bi-LSTM Forecasting Model Integrated with GIS for Spatiotemporal Malaria Endemicity Mapping and Early Warning

Wellie Sulistijanti[1,*], Safaat Yulianto[2], Abdul Syukur[3], Ngatimin[4]

[1,2,4]*Institut Teknologi Statistika dan Bisnis Muhammadyah Semarang, Prof. Hamka Street, Semarang, 50185, Indonesia*

[3]*Universitas Dian Nuswantoro, Imam Bonjol Street, Semarang, 50131, Indonesia*

(Received: February 5, 2026; Revised: April 10, 2026; Accepted: July 10, 2026; Available online: August 18, 2026)

Abstract

Malaria continues to pose a major public health challenge in Indonesia, particularly across eastern provinces where transmission dynamics are shaped by climatic variability, geographic heterogeneity, and historical incidence trends. This study proposes an integrated spatio-temporal malaria forecasting and early warning framework combining Bidirectional Long Short-Term Memory (Bi-LSTM), Geographic Information Systems (GIS), and SHapley Additive exPlanations (SHAP). Monthly malaria incidence and climate data from 12 endemic provinces spanning January 2014 to December 2025 were used, with model training conducted on January 2014 to July 2023 observations and out-of-sample forecasts generated for January 2024 to December 2025. Preprocessing included incidence-rate conversion, province-specific outlier handling, logarithmic transformation, Min-Max normalization, and six-month sliding window segmentation. An ablation study across four architectural configurations confirmed that bidirectional temporal processing was the primary driver of performance improvement, while province embeddings provided no additional benefit given the available training data. The final Bi-LSTM model was benchmarked against Naive Forecasting, SARIMA, and Simple LSTM baselines using RMSE, sMAPE, and R^2 metrics, achieving superior global performance with an RMSE of 0.0496, sMAPE of 14.63%, and R^2 of 0.9597. Province-level evaluation confirmed strong predictive accuracy across most regions, including high-burden areas such as Papua, while reduced performance in West Nusa Tenggara and North Sulawesi was attributable to near-elimination conditions and post-2020 structural breaks in transmission patterns, respectively. SHAP analysis identified historical incidence as the dominant predictor, with rainfall representing the most influential climatic variable. GIS integration translated model outputs into dynamic endemicity maps and a province-level early warning system. These findings demonstrate that the proposed framework supports interpretable malaria forecasting and targeted public health surveillance across geographically heterogeneous endemic regions.

Keywords: Bi-LSTM, Geographic Information Systems, Malaria Prediction, SHAP, Spatiotemporal Analysis

1. Introduction

Malaria remains a major global infectious disease and continues to pose a serious public health burden. In 2024, it is that there would be 282 million cases of malaria with 610,000 deaths worldwide, an increase from 263 million cases with 597,000 deaths in 2023 [1], [2]. This disease is also endemic in tropical regions like Indonesia, contributing 25.8% of estimated malaria cases in the Western Pacific Region [2]. Transmission is heavily affected by geographic and climatic factors, including temperature, rainfall, and humidity, which affect the survival, reproduction, and distribution of *Anopheles* mosquitoes [3]. Temperatures between 25°C to 35°C can accelerate egg laying and shorten the life cycle of *Anopheles* mosquitoes [4], [5]. Nonetheless, elevated temperatures can delay mosquito reproduction and reduce transmission rates [6], [7]. Relative humidity between 55% and 85% is also considered optimal for malaria transmission [6], [8]. Rainfall further contributes to malaria transmission [9].

Since the enactment of Minister of Health Regulation Number 22 of 2022 concerning malaria eradication, malaria cases in Indonesia increased from 443,530 to 735,154 throughout 2022 to 2025 [10]. Despite national malaria eradication efforts, a number of regions in eastern Indonesia still face the problem of this disease, such as Papua, West Papua, East Nusa Tenggara, and West Kalimantan. This growth is affected by population mobility, restricted health services, inconsistent control coverage, and the enhancement of case detection efforts [2]. Climatic variables

***Corresponding author: Wellie Sulistijanti (wellie.sulistijanti@itesa.ac.id)**
DOI: https://doi.org/10.47738/jads.v7i3.1483

additionally influence transmission patterns. Eastern Indonesia experiences continuous malaria transmission throughout the year, attributed to consistent rainfall, with annual precipitation above 3,000 mm and exceeding 6,600 mm in certain parts of Papua and West Papua in 2025 [11]. Conversely, the western and central regions typically have seasonal transmission increases during the wet season [12]. These differences in climate and geography create differences in the number of malaria cases in various provinces, indicating the need for forecasting models that also include spatial and temporal considerations.

Many recent studies have utilized machine learning for malaria forecasting, but most have focused on studying temporal patterns, rather than spatial information [13], [14], [15]. Conversely, GIS-based studies are useful for identifying areas at risk of malaria but are unable to represent changes in transmission over time [15]. However, there is still no integrated framework that combines both spatial and temporal aspects, particularly in countries with diverse malaria patterns such as Indonesia. Despite Bi-LSTM's enhanced temporal modeling capabilities, its opaque nature restricts interpretability. Moreover, research integrating deep learning, GIS, and explainable AI methodologies like SHAP into a unified framework for malaria forecasting remains limited, highlighting a research gap worthy of exploration.

This study formulates a spatio-temporal malaria case prediction system leveraging a combination of Bi-LSTM, GIS, and SHAP to address these constraints. Bi-LSTM processes sequential information in both forward and backward directions, capturing complex malaria fluctuations across 12 endemic provinces up to 2025. A GIS component is systematically implemented to translate model outputs into localized, dynamic risk and endemicity maps. By embedding SHAP into the pipeline, this work systematically resolves model opacity, uncovers spatial heterogeneity in climate drivers, and provides a rigorous epidemiological evaluation. Therefore, this research explicitly seeks to bridge the gap between abstract mathematical modeling and field-level epidemic response, with the ultimate goal of providing a transparent, highly interpretable baseline for national early warning systems and strategic resource allocation in Indonesia.

2. Literature Review

2.1. Temporal Prediction Models for Malaria

Research on malaria prediction has developed rapidly over the past decade, involving diverse methodological approaches. Statistical techniques like ARIMA have been utilized to model seasonal malaria patterns [15], whereas machine learning methodologies such as random forest (RF), boosted regression trees (BRT), and maximal entropy (Maxent) have been employed to model environmental risk factors [14]. The analytical hierarchy process (AHP) and linear regression (LR) have been utilized for less complex predicting applications [13]. Agent-based models have been applied to mimic intricate malaria transmission dynamics [16].

Deep learning methodologies, especially Long Short-Term Memory (LSTM) networks, have become prominent for identifying non-linear temporal patterns in disease prediction [7], [17]. LSTM's ability to retain long-term dependencies makes it more suitable than ARIMA for modeling complex climate-disease relationships. However, standard LSTM processes sequential data unidirectionally from past to present which may limit its sensitivity to fluctuating patterns influenced by both antecedent and subsequent time steps [18].

Bi-LSTM overcomes this weakness by analyzing temporal input in both forward (past-to-future) and backward (future-to-past) directions, enhancing the model's sensitivity to non-linear and fluctuating patterns [17]. Numerous studies have shown that Bi-LSTM attains greater predictive accuracy than standard LSTM, Holt–Winters, Elastic Net, ARIMA, Random Forest, Neural Network, and SVR in diverse disease prediction scenarios and environmental parameter forecasting tasks characterized by dynamic patterns [18], [19], [20]. Its strong performance is mainly supported by its ability to learn complex non-linear dynamic data, which results in lower Mean Absolute Error (MAE) And Root Mean Square Error (RMSE) than models based on ARIMA [21]. Despite these advantages, Bi-LSTM applications in malaria forecasting remain largely confined to studies from a single region, with no published implementation covering the heterogeneous endemic provinces of Eastern Indonesia.

2.2. Spatial Approaches and GIS in Malaria Surveillance

Geographic Information Systems (GIS) have been widely applied to identify spatial distribution and risk patterns of malaria transmission across various geographical settings. Diriba et al. [22] utilized remote sensing and GIS to delineate malaria risk zones in Nekemte City, West Ethiopia, illustrating that environmental spatial data can proficiently identify high-risk locations at the local level. Demoze et al. [8] employed GIS to evaluate the spatial, temporal, and spatio-temporal variations of malaria cases in Southwest Ethiopia, demonstrating that transmission intensity is unevenly distributed and demonstrates significant geographical clustering influenced by local ecological factors. Kitawa and Asfaw [4] enhanced this methodology by using space-time malaria risk mapping in Southern Ethiopia, demonstrating that the incorporation of temporal variation into spatial analysis significantly enhances the identification of transmission hotspots.

In Indonesia, Rejeki et al. [23] used GIS to conduct a spatiotemporal analysis of malaria transmission in Java, revealing that spatial clustering patterns were significantly linked to environmental and demographic parameters. Mwangungulu et al. [24] employed a geospatial model in Tanzania to forecast malaria transmission at the regional level, demonstrating that geospatial methodologies facilitate more focused surveillance by pinpointing provinces and districts with consistently high risk. Byrne et al. [25] employed geostatistical techniques in Laos to delineate the spatial distribution of malaria in in environments with low malaria transmission, whereas Salahi-Moghaddam et al. [26] amalgamated Growing Degree Days (GDD) and MaxEnt spatial predictions to evaluate environmental risk and spatio-temporal malaria transmission in Iran. These studies collectively illustrate that GIS frameworks offer crucial spatial context for malaria surveillance; however, they are constrained in modeling the temporal dynamics that influence transmission cycles.

2.3. Bi-LSTM Integrated with GIS for Malaria Forecasting

While temporal and spatial approaches have each demonstrated utility, studies integrating predictive modelling and GIS for malaria forecasting remain limited. Javaid et al. [27] combined machine learning and GIS for malaria prediction in real time in Punjab, Pakistan, achieving 93.97% accuracy with Random Forest and demonstrating geographically explicit risk mapping. Similarly, Liu et al. [14] applied RF, BRT, and Maxent models using climate and land use data across Southeast Asia and the Western Pacific, obtaining AUC values of 0.994, 0.991, and 0.987, respectively, highlighting the predictive strength of spatial environmental variables. Komura and Matsuoka [13] combined a climate-based model with Analytical Hierarchy Process (AHP) and Linear Regression (LR) to estimate malaria risk in South Kivu, Congo, providing multi-dimensional risk assessment outputs covering risk estimation, model performance, and projected case counts.

However, these integrated approaches share a common limitation: machine learning and statistical models such as RF, BRT, AHP, and LR do not explicitly model sequential temporal dependencies within time series data. They treat climate and spatial covariates as static inputs, without explicitly capturing temporal dependencies. In contrast, Bi-LSTM learns bidirectional temporal patterns, enabling the modeling of non-linear and seasonal malaria dynamics [26]. When integrated with GIS and endemicity-based classification, Bi-LSTM can generate dynamic risk maps, support early warning systems, and guide targeted interventions. This makes Bi-LSTM–GIS integration a promising framework for spatiotemporal malaria forecasting in Indonesia, where transmission patterns vary substantially across provinces.

2.4. Explainability in Deep Learning: SHAP

Despite their high predictive performance, Bi-LSTM models often lack interpretability, providing limited insight into how climatic factors influence malaria outbreaks. To bridge this gap, Explainable AI (XAI) methods such as SHAP are increasingly incorporated into prediction frameworks. According to a recent systematic literature review by Alkhanbouli et al. [28], post-hoc interpretation methods, particularly SHAP, are now recognized as a mandatory framework for transforming complex black-box algorithms into reliable and validated prediction tools. Additionally, research by Ghosh and Khandoker [29] showed that SHAP can explain feature contributions through intuitive visualizations, enhancing model transparency and accountability for medical stakeholders. By combining temporal Bi-LSTM, GIS spatial layout, and regional SHAP values, the predictive model can be decomposed into a transparent and easily interpretable architecture capable of explaining local transmission anomalies and structural shifts across regions.

3. Methodology

This study presents a spatio-temporal malaria prediction model by integrating a Bidirectional Long Short-Term Memory (Bi-LSTM) network with post-hoc SHAP (SHapley Additive exPlanations) analysis and a Geographic Information System. The framework is designed to predict regional malaria incidence, map spatial risks for endemicity assessment, and support public health early warning systems. The methodological workflow consists of five main stages: (1) data collection, (2) data preprocessing, (3) prediction modeling using Bi-LSTM, (4) model evaluation and feature interpretation using SHAP, and (5) malaria endemicity mapping using GIS. The details of the system development flow in this study are illustrated in figure 1.

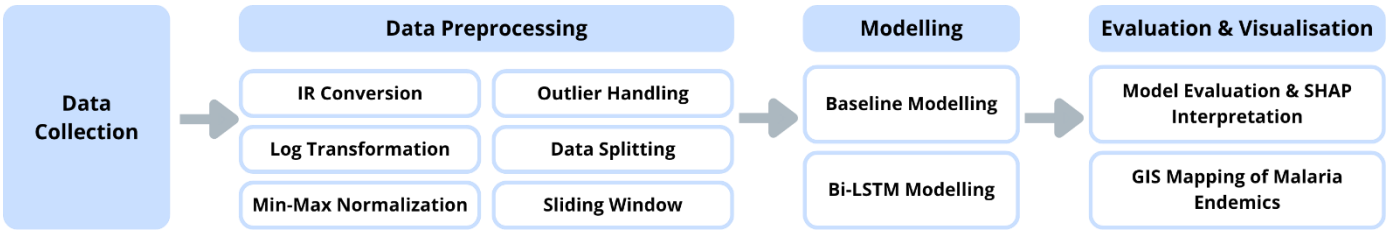

Figure 1. Spatio-temporal Malaria Prediction Model Development Workflow

3.1. Study Area and Data Collection

This study focuses on 12 endemic provinces in Eastern Indonesia: Papua, West Papua, East Nusa Tenggara, West Nusa Tenggara, South Sulawesi, West Sulawesi, Southeast Sulawesi, Central Sulawesi, North Sulawesi, Gorontalo, Maluku, and North Maluku. These provinces were selected based on their consistently high malaria endemicity, collectively accounting for over 95% of Indonesia's total malaria burden throughout the study period.

Monthly malaria case data, climate variables, and provincial population data were compiled from the annual *Province in Figures* (*Provinsi Dalam Angka*) statistical publications released by Statistics Indonesia (Badan Pusat Statistik, BPS) for each of the 12 provinces, covering the period January 2014 to December 2025. Provincial administrative boundary data were obtained from GADM version 4.1 and were used for spatial visualization of forecasting results using GIS. The combined dataset comprised 1,728 monthly observations (12 provinces × 144 months). A complete summary of the variables and their corresponding data sources is presented in table 1.

Table 1. Variables for Spatio-Temporal Malaria Prediction

No	Indicator	Variable	Unit	References
1.	Malaria Transmission	Number of malaria cases per region	cases	[4], [27]
		Historical monthly malaria case data	cases/month	[5], [8], [27], [30]
2.	Environment Factor	Average monthly temperature	°C	[4], [5], [16], [30]
		Monthly precipitation	mm	[4], [5], [22], [30]
		Monthly relative humidity	%	[5], [16], [30]
3.	Spatial Data	Spatial dataset (shapefile)	-	[4], [8], [16], [22]

Although malaria transmission is also influenced by factors such as land use, elevation, healthcare accessibility, population mobility, mosquito density, and socioeconomic conditions, these variables were not included in the present study for two reasons. First, this study intentionally focused on climatic predictors that are publicly accessible from Statistics Indonesia (BPS), enabling transparent data collection and facilitating reproducibility. Second, incorporating numerous heterogeneous predictors would increase model complexity and potentially obscure the interpretation of the SHAP analysis, whereas the objective of this study was to quantify the relative contributions of the principal climatic variables alongside historical malaria incidence.

3.2. Data Preprocessing

Data pre-processing was performed to ensure the data fit the model. First, the number of malaria cases was converted to a monthly incidence rate per 100,000 population using Equation (1):

$$IR = \frac{malaria\_cases}{population} \times 100{,}000 \quad (1)$$

This conversion removed the bias of raw case counts based on population size, which would otherwise over-represent highly populated provinces. This transformation to incidence rates puts all provinces on an epidemiological baseline, ensuring the model reflects the intensity of transmission, not the absolute burden.

Outlier handling was performed using by applying the Interquartile Range (IQR) capping method independently to each province, with bounds defined as $Upper = Q3 + 1.5 \times IQR$ and $Lower = max(Q1 - 1.5 \times IQR, 0)$, where Q_1 and Q_3 denote the first and third quartiles of each province's incidence rate distribution. Of the 1,728-monthly data, 136 values (7.9%) were corrected by this procedure, mostly in provinces with high temporal variability. The incidence rate and rainfall variables were then log-transformed $y = ln\,(x + 1)$ to reduce skewness and stabilize variance while accommodating zero values. However, temperature and humidity were left in their original form because they approximately have symmetric distributions.

To preserve the temporal order of the time series, the dataset consisting of 144 months of observations (January 2014 to December 2025) was chronologically split into 80% training (January 2014 to July 2023; 115 months) and 20% testing (August 2023 to December 2025; 29 months). For transparency, the testing period was further divided into two phases: August to December 2023 for evaluation against observed malaria incidence, and January 2024 to December 2025 for pure forward forecasting. During the 2024–2025 horizon, the model recursively generated malaria incidence predictions using observed climatic variables and its own previous predictions; therefore, all results for this period represent forecasts rather than observed epidemiological data. Finally, all input features were normalized to the [0, 1] range using Min-Max normalization, as defined in Equation (2).

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (2)$$

Normalization was performed globally across all provinces rather than independently for each province to preserve the relative differences in transmission intensity between provinces with high and low transmission burdens. The normalized time-series data was then segmented into sliding windows of six months, with each sequence of six consecutive observations used to predict the incidence rate for the following month. The window length was chosen to capture delayed climatic effects on malaria transmission. A window of six months was selected based on previous malaria forecasting studies reporting delayed climatic effects over several months, as well as preliminary exploratory experiments showing more stable model convergence than shorter windows. Segmentation performed separately for each province produced 1,296 training sequences and 288 testing sequences.

3.3. Bi-LSTM Model Architecture

The proposed model employs a two-layer Bidirectional Long Short-Term Memory (Bi-LSTM) architecture. The overall architecture of the proposed model is illustrated in figure 2.

Each input sample consists of a sequence spanning six months and containing four normalized features (historical malaria incidence, rainfall, humidity, and temperature). These sequences are processed by Bi-LSTM with a two layers and 32 hidden units in each direction. Rather than using the final hidden state returned by the recurrent network, the model extracts the output representation from the final time step as the temporal feature vector. Since the network comprises two recurrent layers, the built-in PyTorch LSTM dropout is applied between recurrent layers, followed by an additional dropout layer (rate = 0.3) before the fully connected layer to further reduce overfitting. Finally, a Softplus activation function (β = 10), as defined in Equation (3), is applied to ensure nonnegative malaria incidence predictions.

ISSN 2723-6471

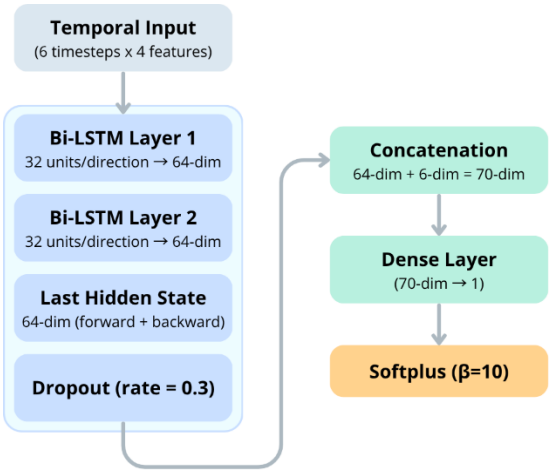

Figure 2. Bi-LSTM Model Architecture

$$f(x) = \frac{1}{\beta} ln(1 + e^{\beta x}) \quad (3)$$

Softplus was selected to ensure non-negative predictions while avoiding the zero-gradient issue of ReLU. The $\beta = 10$ parameter was chosen to make the Softplus function closely approximate ReLU within the normalized incidence range (0–1) while maintaining smooth differentiability near zero.

The model was trained using a combined loss function that incorporates Mean Squared Error (MSE) and Mean Absolute Error (MAE), as described in Equation (4).

$$L = \alpha \cdot MSE + (1 - \alpha) \cdot MAE, \alpha = 0.7 \quad (4)$$

The combined loss function assigns a weight of $\alpha = 0.7$ to the MSE component and $1 - \alpha = 0.3$ to the MAE component, placing greater emphasis on reducing large prediction errors while retaining the robustness of MAE. This formulation mitigates the tendency of optimization based solely on MSE to pull predictions toward the dataset mean, which often results in underestimates the variability observed in provinces with highly fluctuating malaria incidence. By incorporating MAE, the model is better able to preserve extreme incidence patterns without sacrificing sensitivity to substantial forecasting errors.

The model was trained using the Adam optimizer with an initial learning rate of 0.001 and a batch size of 16, for a maximum of 200 epochs. A ReduceLROnPlateau scheduler was employed, reducing the learning rate by a factor of 0.5 when the validation loss failed to improve for 7 consecutive epochs. Early stopping with a patience of 15 epochs was applied based on validation loss, and the model weights corresponding to the lowest validation loss were retained for evaluation. The validation set was constructed chronologically, comprising the final 12 months of each province's training sequences, while the remaining sequences were used for training and shuffled at each epoch. All experiments were implemented in Python 3 using the PyTorch framework and executed on Google Colab, with a fixed random seed (42) applied across NumPy and PyTorch to ensure reproducibility.

3.4. Baseline Models

To contextualize the performance of the proposed Bi-LSTM model, three baseline models were implemented: Naive Forecasting, Seasonal ARIMA (SARIMA), and a unidirectional Simple LSTM. Naive Forecasting served as a lower-bound benchmark by using the previous month's incidence rate as the prediction. SARIMA was fitted independently for each province using logarithmically transformed incidence data. A uniform SARIMA $(1,1,1)(1,1,1,)^{12}$ was intentionally adopted as a conventional seasonal statistical benchmark and applied consistently across all provinces to ensure a standardized comparison with the proposed Bi-LSTM model, rather than individually optimizing province-specific SARIMA configurations. When seasonal convergence failed, a fallback $SARIMA\ (1,1,0)(1,1,1)^{12}$ model was employed. Models were trained on data from January 2014 to December 2023 and evaluated using out-of-sample

forecasts over 24 months. The Simple LSTM retained the same architecture as the proposed model but replaced the bidirectional layer with a unidirectional LSTM.

3.5. Evaluation Metrics

Model performance was evaluated using Root Mean Squared Error (RMSE), coefficient of determination (R^2), and Symmetric Mean Absolute Percentage Error (sMAPE) [7], [31], [32]. In these equations, n represents the total number of observations, y_i denotes the actual malaria value, $\hat{y}_i$ represents the predicted value, and $\bar{y}$ denotes the mean of the actual values. Table 2 summarizes these metrics.

Table 2. Evaluation Metrics for Model Performance Assessment

Metric	Formula	Description
RMSE	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$	Measures average prediction error magnitude; penalizes large deviations [33]
R^2	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$	Proportion of variance in malaria cases explained by the model [33]
sMAPE	$sMAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{\lvert\hat{y}_t-y_t\rvert}{(\lvert\hat{y}_t\rvert+\lvert y_t\rvert)/2}\times 100$	Symmetric relative error; robust for near zero malaria incidence values [32]

3.6. SHAP Interpretability Analysis

To address the limited interpretability of the Bi-LSTM model, post-hoc feature attribution was performed using SHAP (SHapley Additive exPlanations). Given that the model accepts two inputs (temporal sequence and province embedding), a Kernel Explainer was employed, as it is model-agnostic and accommodates arbitrary input structures without requiring direct access to model gradients. SHAP was applied directly to the flattened temporal input of dimension 6×4=24, representing six time steps across four input features.

A background dataset of 100 randomly sampled training sequences was used to establish the reference distribution, and SHAP values were computed for 50 randomly sampled testing sequences using 200 perturbation samples per explanation. SHAP values were averaged across the six time steps to obtain feature attributions, enabling global feature importance ranking and comparison of feature contributions across province. Because a fixed average province embedding was used, the reported SHAP attributions do not capture province-specific embedding effects.

3.7. Malaria Endemicity Mapping Using GIS

Model predictions were integrated with provincial boundaries from GADM Level 1 to develop an interactive GIS-based malaria surveillance system. Predicted incidence rates were categorized according to the Indonesian Ministry of Health thresholds: malaria-free (<1), low (1–5), moderate (5–50), and high (>50) per 1,000 population. An early warning module generated three alert levels: ALERT for high endemicity, WARNING for category escalation relative to the previous month, and WATCH for incidence increases exceeding 20% without a category change. The GIS dashboard was developed using the Streamlit framework and deployed publicly, providing choropleth mapping, temporal trend visualization, early warning alerts, and model performance summaries at the provincial level.

4. Results and Discussion

4.1. Descriptive Statistics of Study Variables

The dataset comprised 1,728 monthly observations spanning January 2014 to December 2025, covering 12 endemic provinces of Eastern Indonesia across 144 consecutive months. Table 3 summarizes the descriptive statistics of all study variables prior to preprocessing transformations. The incidence rate showed a highly skewed distribution, with a mean of 86.03 and a standard deviation of 314.56, indicating substantial variability in malaria transmission intensity across provinces and time periods. This large dispersion suggests that malaria burden was not evenly distributed, with several provinces experiencing very high incidence while others recorded relatively low or near-zero cases.

ISSN 2723-6471

Table 3. Descriptive Statistics

	incidence_rate	rainfall	humidity	temperature
mean	86.03	289.9	81.9	27.24
std	314.56	789.68	5.62	1.63
min	0.00	0.07	57	18.06
max	4498.16	1335.3	100	38.3

Rainfall also showed considerable variation, reflecting differences in climatic conditions across the study regions. In contrast, temperature and humidity exhibited narrower ranges, although they remain epidemiologically relevant because both variables can influence mosquito breeding, survival, and malaria transmission dynamics, as shown in figure 3.

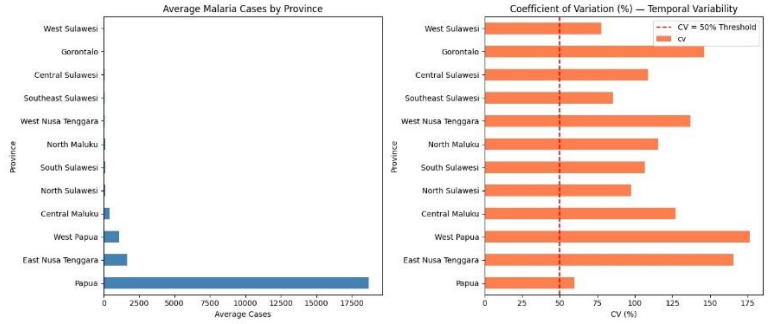

Figure 3. Average malaria raw cases and coefficient of variation across 12 provinces

Figure 3 illustrates substantial disparities in malaria burden across provinces. Papua recorded the highest mean monthly malaria cases at approximately 18,600 cases, followed by East Nusa Tenggara (≈1,650 cases) and West Papua (≈1,050 cases). In contrast, Gorontalo and West Sulawesi exhibited the lowest averages, each below 100 cases per month, representing a difference of more than two orders of magnitude relative to Papua. This pronounced inter-provincial imbalance highlights the epidemiological heterogeneity of the study region and motivated the use of incidence-rate normalization and Min-Max scaling rather than raw case counts in subsequent modeling steps. Furthermore, the coefficient of variation (CV) exceeded 50% in all provinces and surpassed 100% in most cases, with West Papua (CV ≈ 176%) and East Nusa Tenggara (CV ≈ 165%) exhibiting the highest temporal variability, indicating strong seasonal and irregular fluctuations in malaria transmission.

4.2. Preprocessing Outcomes

The preprocessing pipeline substantially improved data consistency and model readiness prior to training. Incidence rates were used to eliminate interprovincial population bias, enabling epidemiological comparisons across regions. Interquartile Range IQR capping was applied separately to each province and modified 136 of the 1,728 observations, primarily in provinces with high temporal variability and known data anomalies. To assess the sensitivity of model performance to outlier treatment, we note that IQR capping modified only 7.9% of observations, predominantly concentrated in provinces with known data anomalies. Rather than removing these observations, the capped values were retained as upper-bound estimates to preserve temporal continuity while limiting the influence of extreme values during model training. Logarithmic transformation further reduced the skewness of incidence rate and rainfall variables, stabilizing variance and reducing the influence of extreme observations. Following chronological train-test splitting, Min-Max normalization standardized all features to a common scale, and segmentation into sliding windows of six months generated 1,296 training sequences and 288 testing sequences.

4.3. Baseline Model Performances

Table 4 shows a comparison of the performance evaluation of the three baseline models, Naive Forecasting, SARIMA, and Simple LSTM.

Table 4. Baseline Model Performance Comparison

No	Model	RMSE	sMAPE	R^2
1.	Naive Forecasting	0.1034	40.17%	0.82
2.	SARIMA	0.1289	45.54%	0.73
3.	Simple LSTM	0.0544	15.47%	0.95

The baseline models exhibited distinct performance characteristics. SARIMA performed the worst (RMSE = 0.1289, sMAPE = 45.54%, R^2 = 0.73), suggesting limited ability to capture the complex malaria dynamics influenced by climatic variability. Naive Forecasting achieved better results (RMSE = 0.1034, R^2 = 0.82), indicating strong short-term temporal dependence in malaria incidence. Simple LSTM delivered the best baseline performance (RMSE = 0.0544, sMAPE = 15.47%, R^2 = 0.95), demonstrating that non-linear temporal relationships play a significant role in malaria transmission and are more effectively captured by deep learning models.

4.4. Ablation Analysis

To evaluate the contribution of each architectural component, an ablation study was conducted across four model configurations: Simple LSTM without province embedding, Simple LSTM with province embedding, Bi-LSTM without province embedding, and the originally proposed Bi-LSTM with province embedding. All configurations shared identical training procedures, hyperparameters, loss functions, and random seeds to ensure fair comparison. The global performance results are summarized in table 5.

Table 5. Ablation Study Results

Model	RMSE	sMAPE	R^2
Simple LSTM (no embedding)	0.0544	15.47%	0.9515
Simple LSTM + embedding	0.0594	22.37%	0.9421
Bi-LSTM (no embedding)	0.0496	14.63%	0.9597
Bi-LSTM + embedding (initial)	0.0544	20.59%	0.9515

4.5. Bi-LSTM Model Performances

The global performance of the proposed Bi-LSTM model on the testing dataset is shown in table 6.

Table 6. Bi-LSTM Global Performance

No	Metric	Value
1.	RMSE	0.0496
2.	sMAPE	14.63%
3.	R^2	0.9597

The proposed Bi-LSTM model achieved a global RMSE of 0.0496, an sMAPE of 14.63%, and an R^2 of 0.9597 on the test set, indicating strong agreement between predicted and observed malaria incidence rates across the 12 study provinces. Compared with the best-performing baseline (Simple LSTM without embedding), the proposed model reduced RMSE from 0.0544 to 0.0496 and improved R^2 from 0.9515 to 0.9597. The sMAPE of 14.63% also represents a substantial improvement over the Simple LSTM baseline (15.47%), SARIMA (45.54%), and Naive Forecasting (40.17%), confirming the superiority of bidirectional temporal learning for capturing complex, nonlinear malaria transmission dynamics across heterogeneous endemic provinces.

The ablation study reported in Section 4.4 further demonstrates that the performance gain is attributable to bidirectional processing rather than to province embeddings. This finding indicates that the historical incidence signal encoded in the input features already carries sufficient province-discriminative information, making explicit province identifiers

ISSN 2723-6471

redundant given the available training data. The final proposed model therefore employs a Bi-LSTM architecture without province embeddings, which achieved the highest global R^2 and lowest RMSE among all four ablation configurations. The training and validation loss curves of the final Bi-LSTM model are presented in figure 4.

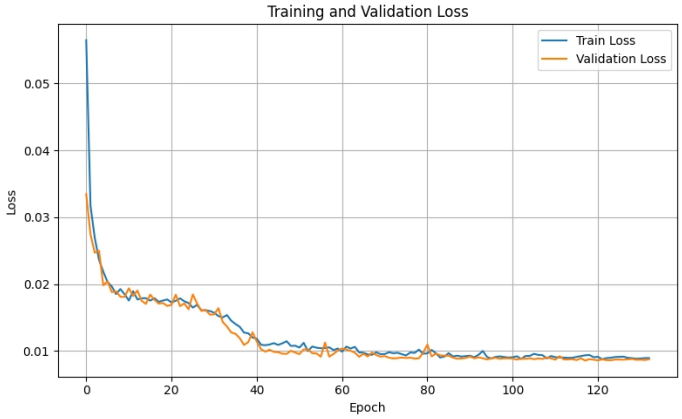

Figure 4. Bi-LSTM training & validation loss

Figure 4 shows the training and validation loss curves of the proposed Bi-LSTM model. Both losses decreased rapidly during the initial epochs before gradually converging to stable values. The close agreement between the training and validation curves, with no observable divergence in later epochs, indicates that the model generalized well to unseen data and did not exhibit substantial overfitting. Furthermore, the smooth convergence behavior suggests that the selected architecture, dropout regularization, and early stopping strategy were effective in achieving stable optimization. To assess spatial robustness, model performance was further evaluated separately for each province, as summarized in table 7.

Table 7. Per-province Evaluation

No	Province	RMSE	sMAPE	R^2
1.	Gorontalo	0.0217	6.05%	0.8061
2.	Maluku	0.0168	5.29%	0.9717
3.	North Maluku	0.0122	5.38%	0.8945
4.	East Nusa Tenggara	0.0398	9.73%	0.8415
5.	West Nusa Tenggara	0.0188	61.84%	0.4721
6.	Papua	0.0596	4.60%	0.9010
7.	West Papua	0.1409	26.02%	0.5547
8.	West Sulawesi	0.0231	10.64%	0.7578
9.	South Sulawesi	0.0124	11.10%	0.9697
10.	Central Sulawesi	0.0151	8.15%	0.9206
11.	Southeast Sulawesi	0.0177	9.42%	0.8978
12.	North Sulawesi	0.0467	19.89%	0.3907

The province-level evaluation in table 7 shows that the proposed Bi-LSTM model achieved strong predictive performance in most provinces, although the magnitude of performance varied across regions. Provinces such as South Sulawesi, Maluku, Central Sulawesi, and Southeast Sulawesi recorded high R^2 values, accompanied by low RMSE and relatively low sMAPE. In addition, the model also demonstrated reasonable generalization ability in high-burden regions such as Papua and West Papua, although prediction errors in these provinces were relatively higher than those observed in lower-burden provinces. These findings indicate that the model was able to capture the temporal variation

ISSN 2723-6471

of malaria incidence rates more accurately in provinces where incidence patterns remained sufficiently observable and relatively consistent during the testing period.

Distinct performance characteristics were observed in West Nusa Tenggara (NTB) and North Sulawesi, where predictive performance was lower than in the other provinces. In NTB, malaria incidence declined substantially following the province's malaria elimination certification, resulting in prolonged near-zero incidence with sparse and irregular fluctuations. Such near-elimination conditions reduce the predictability of the time series and inflate percentage-based error metrics, despite relatively small absolute prediction errors. In North Sulawesi, the model encountered a different challenge due to a marked structural break beginning around 2020, when malaria incidence rapidly decreased from historically high levels to persistently low transmission. This abrupt epidemiological transition introduced inconsistent temporal patterns between the training and testing periods, making it more difficult for the model to learn stable forecasting relationships. These findings suggest that forecasting performance may decrease in provinces undergoing rapid epidemiological transitions, highlighting the potential value of future approaches incorporating regime-switching mechanisms, province-specific fine-tuning, or explicit transition indicators to better capture changing transmission dynamics.

The predominance of the Malaria-Free/Elimination category (90.6% of province-month predictions) reflects the epidemiological reality of Eastern Indonesia during the forecasting period rather than a modeling artifact. This imbalance occurs in the denormalized prediction outputs and does not characterize the training data, where log transformation produced a more balanced target distribution and each of the 12 provinces contributed an equal number of training sequences. Consequently, the model was not biased toward low-incidence provinces and remained capable of accurately forecasting high-transmission regions, as evidenced by the strong predictive performance observed in provinces such as Papua. Therefore, the skewed distribution of predicted endemicity levels primarily reflects the ongoing malaria elimination progress in Eastern Indonesia rather than an imbalance introduced during model training.

4.6. SHAP Interpretation

To improve model interpretability, SHAP analysis was conducted to quantify the contribution of each input variable to the Bi-LSTM predictions. Figure 5 presents the SHAP feature importance and summary plots of the proposed model.

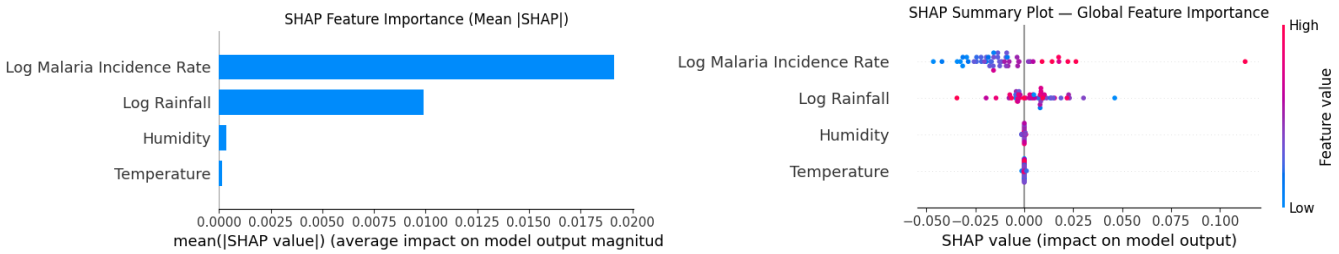

Figure 5. SHAP feature importance & summary plot

Figure 5 shows that historical malaria incidence (Log Malaria Incidence Rate) was the most influential predictor, with a mean absolute SHAP value of approximately 0.019, substantially exceeding all climatic variables. Among the climatic variables, rainfall (Log Rainfall) exhibited the second-highest importance (mean |SHAP| ≈ 0.010), although its contribution remained considerably smaller than that of historical incidence. Humidity and temperature showed only marginal influence, with mean |SHAP| values of approximately 0.0015 and 0.0005 respectively, suggesting that their independent contribution to prediction was minimal in the normalized input space.

4.7. Risk Mapping and Early Warning Using GIS

An interactive GIS dashboard was developed using Streamlit to visualize predicted malaria incidence across the 12 provinces. Figure 6 presents the December 2025 choropleth map, highlighting substantial geographic heterogeneity in malaria transmission intensity across Eastern Indonesia, as shown in figure 6.

 ISSN 2723-6471

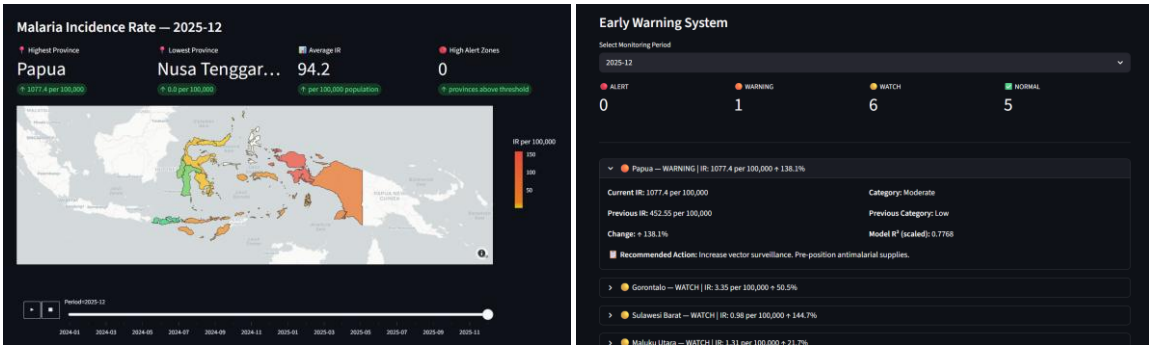

Figure 6. (a) predicted malaria incidence map for December 2025 and (b) province-level early warning interface.

The predicted spatial pattern aligns with known malaria epidemiology in Indonesia. Papua recorded the highest predicted incidence rate at 1,077.4 per 100,000 population in December 2025, remaining within the High endemicity category throughout the prediction horizon, while Papua Barat exhibited moderate-to-high incidence levels. In contrast, the Sulawesi provinces and Maluku archipelago generally showed low-to-moderate incidence, whereas West Nusa Tenggara (NTB) recorded the lowest predicted incidence (0.0 per 100,000), consistent with its near-elimination status since 2019. Across the 24-month prediction period (January 2024 to December 2025), 261 of 288 province-month observations (90.6%) were classified as Malaria-Free, with only 8 and 19 observations falling into the Low and Moderate categories, respectively. To support operational surveillance, the GIS dashboard integrates a province-level early warning module that assigns WATCH, WARNING, or ALERT statuses based on changes in predicted incidence and endemicity categories. This functionality enables health authorities to identify provinces approaching critical transmission thresholds and prioritize timely interventions, resource allocation, and vector control measures.

5. Conclusion

This study proposed an integrated malaria forecasting framework that combines Bidirectional Long Short-Term Memory (Bi-LSTM), Geographic Information Systems (GIS), and SHapley Additive exPlanations (SHAP) to support spatio-temporal malaria surveillance in Indonesia. Using climate variables and historical incidence data from 12 endemic provinces between 2014 and 2025, the proposed model successfully captured both temporal transmission dynamics and spatial heterogeneity. The model achieved strong predictive performance, outperforming Naive Forecasting, SARIMA, and Simple LSTM baselines with an RMSE of 0.0496, sMAPE of 14.63%, and R^2 of 0.9597. These results indicate that the proposed model provides more stable and reliable predictions, as it maintains both low absolute prediction error and controlled relative percentage error on the testing dataset.

Evaluation at the provincial level further showed that the model performed well across most provinces, including high-burden regions such as Papua and West Papua, although prediction errors varied across regions. The model achieved higher reliability in provinces with sufficiently observable and consistent incidence patterns, whereas lower performance was observed in areas experiencing a sharp decline in incidence or approaching malaria elimination, such as West Nusa Tenggara. This finding highlights the importance of evaluation at the provincial level, as global performance metrics alone may not fully reflect local forecasting challenges in regions undergoing major epidemiological transitions. Beyond predictive accuracy, the integration of SHAP improved model transparency by identifying historical incidence as the dominant predictor, followed by rainfall as the most influential climatic factor. The GIS dashboard further translated model outputs into actionable information through endemicity mapping and province-level early warning alerts. Together, these components provide a unified framework capable of supporting data-driven malaria surveillance, risk assessment, and intervention planning in geographically heterogeneous settings.

This study has several limitations. The current model was developed using aggregated data at the provincial level and a fixed input window of six months, which may not fully capture localized transmission dynamics, longer seasonal cycles, or near-zero incidence behavior in elimination settings. In addition, the model was validated only using data from 12 provinces in Eastern Indonesia and has not yet been evaluated in other regions or on independent external datasets, limiting the generalizability of the findings. Future research may explore district level forecasting, adaptive

sequence lengths, zero-inflated or elimination-transition modeling approaches, additional environmental and socioeconomic variables, real-time surveillance integration, and external validation across broader geographic settings to further improve predictive accuracy and operational applicability.

6. Declarations

6.1 Author Contributions

Conceptualization: W.S., S.Y.; Methodology: S.Y., A.S.; Software: N.; Validation: S.Y., N.; Formal Analysis: W.S.; Investigation: W.S.; Resources: S.Y., A.S.; Data Curation: W.S., N.; Writing Original Draft Preparation: W.S.; Writing Review and Editing: S.Y., A.S.; Visualization: N.; All authors have read and agreed to the published version of the manuscript.

6.2 Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3 Funding

This research was supported by Institut Teknologi Statistika dan Bisnis Muhammadiyah Semarang and funded by the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia through the Fundamental Regular Research Grant scheme, Grant Number 082/LL6/AL.04.03/PL-BARU/2026 and A.1/1/062010/LPPM/PL/II/2026

6.4 Institutional Review Board Statement

Not applicable.

6.5 Informed Consent Statement

Not applicable.

6.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] WHO, World Malaria Report. 2023. [Online]. Available: https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2023

[2] W. H. Organization, World malaria report 2025: Addressing the threat of antimalarial drug resistance ISBN. Imperial College London, 2025.

[3] W. Report, World Malaria Report 2021, vol. 13, no. April. 2021.

[4] 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.

[5] 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.

[6] M. Z. Alam, S. M. Niaz Arifin, H. M. Al-Amin, M. S. Alam, and M. S. Rahman, "A spatial agent-based model of Anopheles vagus for malaria epidemiology: Examining the impact of vector control interventions," *Malar. J.*, vol. 16, no. 1, pp. 1–20, 2017, doi: 10.1186/s12936-017-2075-6.

[7] E. Naroum, E. M. Maka, H. Abboubakar, P. Dayang, A. B. Bamana, B. Garga, H. D. Daouda, M. Bakouri, and I. Khan, "Comparative analysis of deep learning and machine learning techniques for forecasting new malaria cases in Cameroon's Adamaoua region," *Intell. Med.*, vol. 11, no. February, pp. 1–9, 2025, doi: 10.1016/j.ibmed.2025.100220.

[8] 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 Heal.*, vol. 12, no. January, pp. 01–11, 2024, doi: 10.3389/fpubh.2024.1466610.

[9] I. Diouf, A. M. Adeola, G. J. Abiodun, C. Lennard, J. M. Shirinde, P. Yaka, J.-A. Ndione, and E. O. Gbobaniyi, "Impact of future climate change on malaria in West Africa," *Theor. Appl. Climatol.*, vol. 147, no. 3–4, pp. 853–865, 2022, doi: 10.1007/s00704-021-03807-6.

[10] Ministry of Health of The Republic of Indonesia, "Malaria Cases in Indonesia," Ministry of Health of The Republic of Indonesia. [Online]. Available: https://malaria.kemkes.go.id/case

[11] BPS-Statistics Indonesia, Stastical Yearbook of Indonesia 2025. BPS-Statistics Indonesia, 2025.

[12] Ministry of Health of The Republic of Indonesia, "National Action Plan for Acceleration of Malaria Elimination 2020-2026 (Revision)," vol. 2026, 2023.

[13] R. Komura and M. Matsuoka, "High-Temporal-Resolution Prediction of Malaria Transmission Risk in South Kivu, Democratic Republic of the Congo, Based on Multi-Criteria Evaluation Using Geospatial Data," I*SPRS Int. J. Geo-Information,* vol. 12, no. 12, pp. 1–19, 2023, doi: 10.3390/ijgi12120489.

[14] X. Liu, C. Song, Z. Ren, and S. Wang, "Predicting the Geographical Distribution of Malaria-Associated Anopheles dirus in the South-East Asia and Western Pacific Regions Under Climate Change Scenarios," *Front. Environ. Sci*., vol. 10, no. June, pp. 1–13, 2022, doi: 10.3389/fenvs.2022.841966.

[15] D. M. Menda, M. Nawa, R. K. Zimba, C. M. Mulikita, J. Mwandia, H. Mwaba, K. Sichinga, "Forecasting Confirmed Malaria Cases in Northwestern Province of Zambia: A Time Series Analysis Using 2014-2020 Routine Data," *Adv. Public Heal.,* vol. 2021, pp. 1–8, 2021, doi: 10.1155/2021/6522352.

[16] 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. October, pp. 1–13, 2019, doi: 10.1016/j.ecoinf.2019.101006.

[17] A. Gaouar, S. Hamza, A. Rahmoun, M. El, and H. Daho, "Informatics in Medicine Unlocked Explainable AI for early malaria detection using stacked-LSTM and attention mechanisms," *Informatics Med. Unlocked,* vol. 57, no. March, pp. 1–16, 2025, doi: 10.1016/j.imu.2025.101667.

[18] L. Muflikhah, N. Widodo, N. Yudistira, and A. Ridok, "Drug Resistant Prediction Based on Plasmodium Falciparum DNA-Barcoding using Bidirectional Long Short Term Memory Method," vol. 14, no. 7, pp. 433–440, 2023, doi: 10.14569/IJACSA.2023.0140747.

[19] T. Fisher, S. Rojas-galeano, and D. Fernandez-reyes, "Specialist hybrid models with asymmetric training for malaria prevalence prediction," *Front. Public Heal.*, vol. 11, no. 1207624, pp. 01–09, 2023, doi: 10.3389/fpubh.2023.1207624.

[20] Chengqi Wang, Yibo Dong, Chang Li, Jenna Oberstaller, Min Zhang, Justin Gibbons, Camilla Valente Pires, Mianli Xiao, Lei Zhu, Rays H. Y. Jiang, Kami Kim, Jun Miao, Thomas D. Otto, Liwang Cui, John H. Adams & Xiaoming Liu, "MalariaSED : a deep learning framework to decipher the regulatory contributions of noncoding variants in malaria parasites," *Genome Biol.,* pp. 1–20, 2023, doi: 10.1186/s13059-023-03063-z.

[21] 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 Explor. Exploit.*, 2024, doi: 10.1177/01445987241269496.

[22] 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, pp. 1–17, 2024, doi: 10.1186/s12942-024-00386-3.

[23] D. S. S. Rejeki, S. Nurlaela, D. Octaviana, B. Wijayanto, and S. Solikhah, "Clusters of malaria cases at sub-district level in endemic area in Java Island, Indonesia," *Geospat. Health,* vol. 17, no. 1, 2022, doi: 10.4081/gh.2022.1048.

[24] S. P. Mwangungulu, D. Dorothea, Z. R. Ngereja, and E. W. Kaindoa, "Geospatial based model for malaria risk prediction in Kilombero valley, South-eastern, Tanzania," *PLoS One*, vol. 18, no. 10 October, pp. 1–23, 2023, doi: 10.1371/journal.pone.0293201.

[25] I. Byrne, E. Cramer, L. Nelli, F. Rerolle, L. Wu, C. Patterson, J. Rosado, E. Dumont, K. K. A. Tetteh, E. Dantzer, B. Hongvanthong, K. M. Fornace, G. Stresman, A. Lover, A. Bennett, and C. Drakeley, "Characterizing the spatial distribution of multiple malaria diagnostic endpoints in a low-transmission setting in Lao PDR," *Front. Med*., vol. 9, pp. 01–13, 2022, doi: 10.3389/fmed.2022.929366.

[26] A. Salahi-Moghaddam, H. Turki, M. Yeryan, and M. V. Fuentes, "Spatio-temporal Prediction of the Malaria Transmission Risk in Minab District (Hormozgan Province, Southern Iran)," *Acta Parasitol.*, vol. 67, no. 4, pp. 1500–1513, 2022, doi: 10.1007/s11686-022-00598-2.

[27] 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, pp. 1–33, 2023, doi: 10.3390/ijerph20043740.

[28] R. Alkhanbouli, H. Matar Abdulla Almadhaani, F. Alhosani, and M. C. E. Simsekler, "The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions.," *BMC Med. Inform. Decis. Mak*., vol. 25, no. 1, pp. 1–17, Mar. 2025, doi: 10.1186/s12911-025-02944-6.

[29] S. K. Ghosh and A. H. Khandoker, "Investigation on explainable machine learning models to predict chronic kidney diseases," *Sci. Rep.*, vol. 14, no. 1, pp. 1–15, 2024, doi: 10.1038/s41598-024-54375-4.

[30] W. Haileselassie, A. Ejigu, T. Alemu, S. Workneh, M. Habtemichael, R. E. David, K. Lelisa, W. Deressa, G. Yan, D. M. Parker, and B. Taye, "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.

[31] A. Fayyazbakhsh, T. Kienberger, and J. Vopava-wrienz, "Comparative Analysis of Load Profile Forecasting: LSTM, SVR, and Ensemble Approaches for Singular and Cumulative Load Categories," *Smart Cities*, vol. 8, no. 65, pp. 1–28, 2025, doi: 10.3390/smartcities8020065

[32] Y. Xu, B. Guo, Y. Chen, and X. Liu, "Multiscale Nonlinear Forecasting of Government Bond Yields and Volatility via a Hybrid VMD – LSTM Framework," *Mathematics*, vol. 14, no. 1297, pp. 1–22, 2026, doi: 10.3390/math14081297.

[33] W. H. Greene, Econometric Analysis, 8th ed. Pearson, 2020.