



Prediction of Malaria Incidence using Machine Learning Models in the Presence of Climatic Factors in Kebbi State, Nigeria

Hadiza Sai'd Madawaki¹, Yakubu Musa², Umar Usman³ & Abdulkarim Bello⁴

^{1,2,3}Department of Statistics, Usmanu Danfodiyo University Sokoto, Nigeria

⁴Department of Computer Science Usmanu Danfodiyo University Sokoto Nigeria

*Corresponding Author's Email: hadizamadawaki19@gmail.com

Received on: August, 2026

Revised and Accepted on: September, 2026

Published on: October, 2026

ABSTRACT

Malaria remains a critical public health challenge in Nigeria, with Kebbi State experiencing significant transmission burdens influenced by climatic variability. Traditional statistical methods, while useful for identifying associations, often fail to capture the complex, nonlinear, and delayed effects of climatic variables on malaria transmission. This study aimed to assess the impact of key climatic variables on malaria incidence and develop predictive models using machine learning approaches, specifically Random Forest (RF) and Artificial Neural Networks (ANN), in Kebbi State, Nigeria. Monthly malaria incidence data from 2014 to 2022 across all 21 Local Government Areas in Kebbi State were combined with corresponding meteorological records from the Nigerian Meteorological Agency. Preprocessing included normalization, temporal feature engineering (lagged and rolling means), and creation of both regression and binary classification targets. Model performance was evaluated on an independent test set (20% of data) using R^2 , RMSE, MAE for regression, and accuracy, precision, recall, and F1-score for classification. RF regression achieved $R^2 = 0.997$, RMSE = 110.93, and MAE = 37.52 on the test set, while ANN achieved $R^2 = 0.991$, RMSE = 178.43, and MAE = 84.57. For classification, RF achieved 89.5% test set accuracy with balanced precision and recall, whereas ANN achieved 89.1% test set accuracy. Feature importance analysis confirmed lagged rainfall and humidity as the most influential predictors. Traditional statistical models (MLR) showed limited explanatory power ($R^2 = 0.068-0.097$), justifying the machine learning approach. Machine learning models, particularly Random Forest, effectively capture the nonlinear, time-dependent relationships between climate and malaria transmission, significantly outperforming conventional statistical approaches. This study provides a robust predictive framework for malaria surveillance and early warning systems in Kebbi State, highlighting the importance of integrating climatic and temporal predictors for improved disease forecasting.

Keywords: Malaria incidence, Climatic factors, Random Forest, Artificial Neural Networks, Machine Learning, Predictive modeling, Epidemiology

1.0 INTRODUCTION

Malaria remains a significant public health burden in sub-Saharan Africa, with Nigeria bearing the highest global share of cases and deaths. According to the 2024 World Malaria Report, Nigeria accounted for 25.9% of global malaria cases and 30.9% of related deaths in 2023, with approximately 97% of its population at risk (WHO, 2024). Within Nigeria, Kebbi State, located in the northwestern region, is particularly vulnerable due to its tropical savanna climate, characterized by high temperatures, distinct wet and dry seasons, and variations in humidity all of which create conducive conditions for *Anopheles* mosquito breeding and malaria transmission (Köppen-Geiger classification). The relationship between climatic factors (temperature, rainfall, relative humidity, and wind) and malaria incidence is well-established, as these variables directly influence the lifecycle, reproduction, and survival of mosquito vectors, as well as the development rate of the *Plasmodium* parasite (Mordecai et al., 2013; Beck-Johnson et al., 2013).

Traditional statistical models, such as linear regression, generalized linear mixed models, and cluster-

integrated regression, have been extensively employed to quantify the association between meteorological variables and malaria incidence (Amadi & Erandi, 2024; Ogunsakin et al., 2024; Padhi et al., 2024; Colonia et al., 2024). While these approaches offer valuable inferential insights, they are constrained by several limitations: they often assume linear or log-linear relationships, struggle to capture complex nonlinear interactions and delayed climatic effects, and typically yield modest predictive power (low R^2 values). Moreover, many studies apply these methods without accommodating the localized, heterogeneous transmission dynamics observed across different regions, such as the Local Government Areas within Kebbi State.

In recent years, machine learning (ML) techniques have emerged as powerful alternatives for modeling complex, nonlinear relationships in epidemiological data. Among these, Artificial Neural Networks (ANN) and Random Forest (RF) have shown particular promise for malaria prediction tasks, given their ability to handle high-dimensional data, identify feature importance, and model temporal dependencies without

rigid parametric assumptions (Okundalaye et al., 2025; Adamu et al., 2021; Olatayo et al., 2025). ANN, which underpin deep learning, excel at capturing intricate patterns in time-series data, while RF, an ensemble learning method, provides robustness against overfitting and offers interpretable measures of variable importance. Despite this growing body of evidence, there is a notable absence of localized, systematic comparative evaluations of these two leading ML models for malaria incidence prediction in Kebbi State. To date, no study has rigorously contrasted the predictive performance of ANN and RF using the same climatic dataset and a consistent evaluation framework within this specific geographical and epidemiological context.

The aim of this study is to analyze the efficient prediction of malaria incidence in Kebbi State as a function of climatic factors, using Random Forest and Artificial Neural Network models. Specifically, the objectives are to: (a) identify the climatic factors most strongly associated with malaria incidence in Kebbi

2.0 MATERIALS AND METHODS

2.1 Study Area and Design

This study was conducted in Kebbi State, Nigeria, located at an elevation of 156.16 meters above sea level with a tropical savanna climate. The state experiences a distinct wet season from May to October and a dry season from November to April. Monthly data spanning January 2014 to December 2022 were utilized, covering all 21 Local Government Areas (LGAs).

2.1.1 Data Sources

Two primary datasets were utilized:

- i. **Malaria Incidence Data:** Monthly confirmed malaria case counts obtained from the Health Management Information System (HMIS), Directorate of Planning and Statistics, Ministry of Health, Kebbi State.
- ii. **Climatic Data:** Monthly aggregated meteorological measurements sourced from the Nigerian Meteorological Agency (NiMet), including:
 - a) Rainfall (mm)
 - b) Temperature (minimum and maximum, °C)
 - c) Relative humidity (%)
 - d) Wind speed (m/s)

These variables were selected based on extensive epidemiological literature demonstrating their influence on mosquito breeding habitats, parasite development cycles, and vector-host interaction patterns.

State using robust analytical techniques; (b) build predictive ML models using ANN and RF; and (c) compare the forecast performance of ANN and RF models.

The contribution to knowledge of this work is twofold. First, it provides the first localized, empirical comparison of Random Forest and Artificial Neural Network models for malaria incidence prediction in Kebbi State, thereby filling a critical gap in the literature on climate-sensitive disease modeling in northwestern Nigeria. Second, by identifying the key climatic drivers and evaluating model performance within a unified framework, this study offers actionable insights for enhancing malaria early warning systems, supporting data-driven public health decision-making, and enabling timely, targeted interventions in a region highly susceptible to climate-induced malaria outbreaks. Given the accelerating impacts of climate change, this research also contributes to the broader discourse on adaptive disease management strategies in endemic settings.

2.1.2 Data Preprocessing and Feature Engineering

Data preprocessing was undertaken to ensure reliability, continuity, and compatibility between datasets:

2.1.3 Missing Data Handling: Linear interpolation was applied for missing values embedded within continuous monthly intervals, and mean substitution was used for isolated missing points.

2.1.4 Outlier Detection: The Interquartile Range (IQR) technique was employed to identify and correct anomalous values.

2.1.5 Temporal Synchronization: Climatic variables were aggregated to monthly averages or totals to align temporally with malaria incidence reporting cycles.

2.1.6 Feature Scaling: Min-Max normalization was implemented for ANN inputs to improve learning stability:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (3.1)$$

Where X represents the original feature value, X-min and X-max are the minimum and maximum values of that feature across the dataset, and X-norm is the normalized value scaled to the range [0,1]. This transformation was applied to all input features for the ANN model to improve gradient-based learning stability. Random Forest models, being scale-invariant, received no normalization.

2.1.7 Feature Engineering: Temporal dependencies were captured through the creation of:

- i. Lagged variables (lag1, lag2) for both malaria and climatic variables
- ii. Rolling 3-month means for malaria-related variables
- iii. One-month lags for climatic variables

2.1.8 Target Variable Engineering: Two target variables were created:

- i. **Regression Target:** Continuous monthly confirmed malaria cases (Mean = 2,172; SD = 1,851)
- ii. **Binary Classification Target:** High-incidence months (1) defined as exceeding LGA-specific mean + 1 standard deviation; low-incidence months (0) otherwise.

2.2 Statistical Analysis

2.2.1 Pearson Correlation Analysis was conducted to measure the strength and direction of linear associations between malaria incidence and each climatic factor.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (3.2)$$

Where:

X_i and Y_i are individual data points

$\bar{X}$ and $\bar{Y}$ are means of x and y datasets

Values close to +1 or -1 indicate strong positive or negative relationships, respectively, while values near zero suggest weak linear dependence.

2.2.2 Multiple Linear Regression (MLR) was employed as a benchmark modeling framework:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \quad (3.3)$$

Y is the dependent variable

$X_1, X_2, X_3,$ and X_4 are independent variables

Where β coefficients represent marginal effects

ε denotes the error term.

Parameters were estimated using Ordinary Least Squares (OLS), and model performance was evaluated using R^2 and adjusted R^2 .

2.3 Time-Series Validation Strategy

To address the temporal dependence in the data, a rigorous time-series cross-validation approach was implemented rather than random splitting. The data were ordered chronologically and divided using expanding windows:

2.3.1 Training set: Observations from January 2014 to December 2020 ($n = 1,848$).

2.3.2 Validation set: Observations from January 2021 to December 2021 ($n = 252$).

2.3.3 Test set: Observations from January 2022 to December 2022 ($n = 252$)

This approach ensures that predictions are made for future time periods using only past data, simulating real-world deployment conditions and preventing the data leakage that would occur with random splitting.

2.4 Machine Learning Models

2.4.1 Random Forest

Random Forest (RF) was implemented as an ensemble learning method constructing multiple decision trees using bootstrapped subsets of data and randomly selected features at each split. Feature importance was computed using total decrease in impurity:

$$FI_j \Delta Impurity_j$$

For regression, predictions were the average across all trees:

$$\hat{Y} = \frac{1}{M} \sum_{m=1}^M \hat{Y}_m \quad (3.5)$$

Where M is the number of trees and $\hat{Y}_m$ is the prediction from the m -th tree.

Hyper-parameter optimization was performed using GridSearchCV with five-fold cross-validation, exploring:

- i. $n_estimators$: {100, 200, 300}
- ii. max_depth : {5, 10, 15, 20}
- iii. $max_features$: {0.5, 0.75, $\sqrt{\cdot}$ }
- iv. $min_samples_split$: {2, 4, 6, 8}
- v. $min_samples_leaf$: {1, 2, 3, 5, 7}

2.4.2 Artificial Neural Network

A feed-forward multilayer perceptron (MLP) structure was implemented comprising an input layer, two hidden layers (12 and 8 neurons), and an output layer. The network employed Rectified Linear Unit (ReLU) activation for hidden layers and linear activation for the output layer.

The ANN was trained using supervised learning with backpropagation and gradient descent-based optimization, minimizing Mean Squared Error (MSE):

$$y_i(k) = F(\sum_{i=0}^m w_i(k)x_i(k) + b) \quad (3.6)$$

Where:

$Y_i(k)$ is the output of neuron I at iteration k

F is the activation function

$W_i(k)$ is the weight at iteration k

$X_i(k)$ is the input at iteration k

b is the bias

Hyperparameter optimization was performed using RandomizedSearchCV with three-fold cross-validation, exploring:

- i. Neurons: {16, 32, 64}
- ii. Learning rate: {0.001, 0.005, 0.01}
- iii. Dropout rate: {0.1, 0.2, 0.3}
- iv. Epochs: {50, 100}
- v. Batch size: {16, 32}

2.5 Model Evaluation Metrics

2.5.1 Regression Metrics:

- i. Coefficient of Determination (R^2): Measures proportion of variance explained
- ii. Root Mean Square Error (RMSE): Penalizes larger errors more heavily
- iii. Mean Absolute Error (MAE): Average magnitude of deviation
 - a. $MAE = \frac{1}{n} \sum_{i=1}^n |e_i|$ (3.7)

Where n is the number of observations in the test data set and $e_i = (A - \hat{A})$, where A is the actual value and $\hat{A}$ is the estimated value.

2.5.2 Classification Metrics:

1. Accuracy: Proportion of correctly classified instances

- b. Precision: Proportion of true positive predictions among all positive predictions
- c. Recall: Proportion of true positive predictions among all actual positives
- d. F1-Score: Harmonic mean of precision and recall

All models were trained on 80% of the data and evaluated on the remaining 20% test set.

3.0 RESULTS

3.1. Descriptive Statistics

The final analytical dataset comprised 2,478 monthly observations across all 21 LGAs from 2014 to 2022. Summary statistics revealed considerable variation in both climatic and malaria variables (Table 1).

Table 1: Summary Statistics of Key Variables (2014-2022)

Statistic	Confirmed Cases	Humidity (%)	Rainfall (mm)	Max Temp (°C)	Min Temp (°C)	Temp	Wind Speed (m/s)
Count	2,478	2,478	2,478	2,478	2,478		2,478
Mean	2,172.00	50.00	89.00	35.00	22.00		4.00
SD	1,851.00	20.00	120.00	3.00	4.00		1.00
Minimum	3.00	12.00	0.00	30.00	15.00		1.00
Maximum	16,954.00	83.00	466.00	41.00	28.00		7.00

The mean monthly confirmed malaria cases stood at approximately 2,172 with considerable variation ($SD \approx 1,851$). Rainfall ranged from 0 to 466 mm, and relative humidity varied widely between 12% and 83%, reflecting strong seasonal fluctuations.

3.2.1 Pearson Correlation Analysis

Pearson correlation analysis between monthly confirmed malaria cases and climatic variables revealed negligible linear associations, with all correlation coefficients approximately 0.00 (Table 2).

3.2 Statistical Analysis Results

Table 2: Pearson Correlation between Climatic Variables and Malaria Incidence

Variable	Pearson r	95% CI	p-value
Humidity (%)	0.003	[-0.038, 0.044]	<0.001
Rainfall (mm)	0.001	[-0.040, 0.042]	<0.001
Wind speed (m/s)	-0.002	[-0.043, 0.039]	<0.001
Max temperature	-0.004	[-0.045, 0.037]	<0.001
Min temperature	-0.001	[-0.042, 0.040]	<0.001

These weak correlations indicated that immediate climatic conditions do not exert strong direct linear effects on monthly malaria incidence, supporting the need for nonlinear modeling approaches.

3.2.2 Multiple Linear Regression (OLS)

The multiple linear regression model was statistically significant (F-test $p < 0.001$) but explained only 6.8% of the variance ($R^2 = 0.068$) (Table 3).

Table 3: Multiple Linear Regression Results (OLS)

Variable	Coefficient (β)	Std. Error	p-value	95% CI
Rainfall (mm)	1.46	0.62	0.018	0.24, 2.68
Humidity (%)	23.53	5.06	<0.001	13.61, 33.45
Max temperature	50.70	34.09	0.137	-16.12, 117.52
Min temperature	-133.20	21.75	<0.001	-175.83, -90.57
Wind speed	-160.72	40.23	<0.001	-239.57, -81.87

Humidity ($\beta = 23.53$, $p < 0.001$) and rainfall ($\beta = 1.46$, $p = 0.018$) showed positive and statistically significant associations with malaria incidence. Minimum temperature ($\beta = -133.20$, $p < 0.001$) and wind speed ($\beta = -160.72$, $p < 0.001$) exhibited significant negative effects. Maximum temperature ($\beta = 50.70$, $p = 0.137$) was not statistically significant.

3.3 Lagged Regression Model

Incorporating one-month lagged climatic variables improved explanatory power ($R^2 = 0.097$), confirming that climatic effects operate through delayed mechanisms (Table 4).

Table 4: Lagged Climatic Effects on Malaria Incidence (OLS)

Variable (Lag-1)	Coefficient (β)	Std. Error	p-value	95% CI
Rainfall	2.28	0.61	<0.001	1.08, 3.48
Humidity	23.35	4.99	<0.001	13.57, 33.13
Max temperature	23.74	33.60	0.480	-42.11, 89.59
Min temperature	-147.29	21.50	<0.001	-189.43, -105.15]
Wind speed	-19.72	39.63	0.619	-97.39, 57.95

3.4 Machine Learning Model Performance

3.4.1 Random Forest Model Performance

Following systematic hyperparameter optimization, the optimal Random Forest configurations were identified (Table 5).

Table 5: Optimal Random Forest Hyperparameters

Hyperparameter	Regression	Classification
n_estimators	100	100
max_depth	15	20
max_features	0.5	0.75
min_samples_split	8	2
min_samples_leaf	3	2
Bootstrap	True	True

Table 6: Final Performance of Optimized Random Forest Models (Time-Series Validation)

Table 6a: Regression Performance

Metric	Training (2014-2020)	Test (2022)
RMSE	94.27	110.93
MAE	30.48	37.52
R ²	0.9981	0.9967

Table 6b: Classification Performance

Metric	Training (2014-2020)	Test (2022)
Accuracy	0.9955	0.8952
Precision	0.9955	0.8882
Recall	0.9955	0.8952
F1-Score	0.9954	0.8901

The RF regression model achieved exceptional performance with $R^2 = 0.9967$, RMSE = 110.93, and MAE = 37.52. The classification model demonstrated

strong generalization with test accuracy of 89.52% and balanced precision, recall, and F1-scores above 0.88.

3.4.2 ANN Model Performance

Optimal ANN hyperparameters identified through randomized search are presented in Table 7.

Table 7: Optimal ANN Hyperparameters

Hyperparameter	Regression	Classification
Neurons (per hidden layer)	64	64
Learning rate	0.005	0.001
Dropout rate	0.2	0.3
Epochs	100	50
Batch size	32	16

Table 8: Final Performance of Optimized ANN Models (Time-Series Validation)

Table 8a: Regression Performance

Metric	Training (2014-2020)	Test (2022)
RMSE	162.15	178.43
MAE	75.32	84.57
R^2	0.9934	0.991

Table 8b: Classification Performance

Metric	Training (2014-2020)	Test (2022)
Accuracy	0.921	0.891
Precision	0.917	0.884
Recall	0.921	0.891
F1-Score	0.916	0.886

The ANN regression model achieved $R^2 = 0.991$, RMSE = 178.43, and MAE = 84.57. The classification model achieved test accuracy of 89.1% with balanced precision, recall, and F1-score above 0.88.

3.5 Comparative Model Analysis

Table 9a: Comparative Summary of Regression Model Performance

Model	RMSE	MAE	R^2
Random Forest	110.93	37.52	0.9967
ANN	178.43	84.57	0.991
MLR (no lag)	1,782.51	1,420.36	0.068
MLR (lag-1)	1,759.23	1,398.71	0.097

Table 9b: Comparative Summary of Classification Model Performance

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.8952	0.8882	0.8952	0.8901
ANN	0.891	0.884	0.891	0.886

The Random Forest model consistently outperformed the ANN in both regression and classification tasks, demonstrating superior predictive accuracy and generalization capability.

3.6 Feature Importance Analysis

Random Forest feature importance analysis confirmed the relevance of lagged climatic predictors (Table 10).

Table 10: Random Forest Feature Importance for Climatic Variables

Variable	Importance Score	Normalized Importance (%)
humidity_pct_lag1	0.1847	18.47
rain_mm_lag1	0.1623	16.23
humidity_pct	0.1251	12.51



rain_mm	0.1189	11.89
temp_min_lag1	0.1042	10.42
temp_min	0.0987	9.87
temp_max_lag1	0.0812	8.12
wind_ms_lag1	0.0734	7.34
temp_max	0.0592	5.92
wind_ms	0.0423	4.23

Note: Importance scores are based on the total decrease in node impurity (MSE reduction) averaged across all trees, normalized to sum to 1.0.

Lagged humidity (18.47%) and lagged rainfall (16.23%) emerged as the most influential predictors, followed by contemporaneous humidity (12.51%) and rainfall (11.89%). Temperature variables showed moderate importance, while wind speed had the lowest predictive influence.

4.0 DISCUSSION

4.1.1 Limitations of Traditional Statistical Models

The study findings revealed that traditional statistical methods, while useful for identifying associations, have limited predictive utility for malaria incidence forecasting. Pearson correlation analysis showed negligible linear associations between climatic variables and malaria incidence ($r \approx 0.00$), with narrow confidence intervals indicating precision but not meaningful association. Multiple linear regression explained only 6.8% of the variance ($R^2 = 0.068$). Even with the incorporation of lagged variables, explanatory power remained modest ($R^2 = 0.097$).

These results align with previous studies that have demonstrated the inadequacy of linear models for capturing malaria transmission dynamics. Nanvyat et al. (2018) similarly found low model deviance explained (35.4%) when using climatic factors alone in Plateau State, Nigeria. The weak linear relationships are epidemiologically expected, as malaria transmission depends on delayed mosquito breeding cycles, parasite incubation, and cumulative environmental exposure rather than instantaneous climatic states.

The low explanatory power confirms that malaria transmission dynamics are complex and nonlinear, justifying the transition to advanced machine learning approaches capable of capturing nonlinear interactions, temporal dependencies, and higher-order effects that classical statistical methods cannot adequately represent.

4.1.2 Superior Performance of Machine Learning Models

The Random Forest model demonstrated exceptional predictive performance, achieving near-perfect $R^2 = 0.9967$ for regression and 89.52% accuracy for classification. The ANN also performed strongly ($R^2 = 0.991$, accuracy = 89.1%), though slightly less effectively than RF.

The superior performance of Random Forest can be attributed to several factors. Its ensemble nature aggregates multiple decision trees, effectively handling nonlinearities, variable interactions, and noise without requiring feature scaling or extensive preprocessing. This is particularly relevant to climatic data characterized by natural variability and irregular temporal behavior. The model's resilience to overfitting through bootstrap aggregation and feature randomness makes it well-suited for datasets of moderate size, though appropriate validation strategies remain essential.

These findings are consistent with previous studies. Adamu and Singh (2021) compared six machine learning classifiers for malaria prediction, finding RF achieved 97.72% accuracy, the highest among all models evaluated. Lee et al. (2021) also found RF to be the most effective model after addressing data imbalance concerns.

The relatively lower performance of ANN in this context may be partly attributed to its sensitivity to dataset size and parameter configuration. ANNs generally require sufficient data volume to learn complex patterns without overfitting, though the exact requirement depends on factors including model architecture, regularization techniques, feature dimensionality, and the signal-to-noise ratio in the data. In this study, the dataset of 2,268 observations may be approaching the lower bound for effective ANN generalization, whereas RF's ensemble approach is more sample-efficient for this type of structured tabular data. As noted by Uzun Ozsahin et al. (2024), while ANN performed well in training, RF demonstrated superior resilience and generalization during testing, consistent with our findings.

4.1.3 Importance of Lagged Climatic Variables

The significance of lagged climatic variables, particularly lagged rainfall and humidity, emerged as a critical finding. The models demonstrated that rainfall in the preceding one to two months had stronger predictive influence than current-month rainfall, supporting the delayed biological response observed in malaria transmission cycles.

This finding is epidemiologically robust. Rainfall creates breeding habitats that result in increased mosquito populations after an incubation period of

approximately two to three weeks. Similarly, relative humidity supports mosquito survival and activity, with effects that accumulate over time. These lagged relationships have been reported in studies conducted in Ghana (Oheneba-Dornyo et al., 2022) and Kenya (Nyawanda et al., 2023), where rainfall lag of one to two months yielded the highest correlation with malaria case counts.

The delayed nature of climatic effects underscores the importance of incorporating temporal dynamics into predictive modeling. Forecasting systems should not rely solely on current meteorological readings but consider prior climatic conditions to anticipate potential malaria outbreaks.

4.4.4 Implications for Public Health Policy

The results demonstrate the feasibility of deploying machine learning-based early warning frameworks in local health systems to guide proactive interventions. The ability of the models, particularly Random Forest, to predict malaria trends with high precision provides a data-driven basis for:

1. **Resource Allocation:** Anticipating high-risk periods allows for strategic distribution of insecticides, antimalarial drugs, and medical personnel.
 - i. **Vector Control Scheduling:** Timing of Indoor Residual Spraying (IRS) and Insecticide-Treated Net (ITN) distribution can be optimized.
 - ii. **Public Health Sensitization:** Communities can be alerted before peak transmission seasons, promoting preventive behaviors.
 - iii. **Health System Preparedness:** Hospitals and clinics can prepare for increased caseloads in predicted outbreak periods.

The demonstrated influence of humidity and rainfall highlights the potential value of integrating meteorological monitoring stations and remote sensing technologies into local health surveillance frameworks. Such integration would facilitate real-time risk assessment, improving preparedness and response efficiency in malaria control programs.

4.1.5 Methodological Considerations

The use of time-series cross-validation in this study represents a significant improvement over random splitting approaches common in the literature. By training on earlier years (2014-2020) and testing on later years (2022), we ensure that predictions are made for future periods using only past data, better approximating real-world deployment scenarios.

However, even with proper validation, the high R^2 values warrant careful interpretation. While the model demonstrates excellent predictive capability for the test period, this does not guarantee equal performance

during all future periods, particularly if climate patterns or malaria control interventions change substantially.

4.1.6 Study Limitations and Future Directions

Several limitations should be acknowledged. First, the study focused exclusively on climatic and temporal predictors, excluding socioeconomic, environmental, and behavioral variables such as vegetation cover, housing conditions, sanitation practices, and population movement that may influence transmission patterns. Incorporating these factors could further improve model generalization and explanatory depth. Second, the temperature range in the study area remained relatively stable within the optimal breeding threshold for *Anopheles* mosquitoes throughout the year, limiting temperature's discriminative power in the models. Future studies in regions with greater temperature variability may reveal stronger temperature effects.

Third, while the Random Forest model demonstrated excellent fit on the test data, slight prediction deviations were observed during extreme climatic periods, suggesting that additional covariates could further enhance predictive stability.

Fourth, although the study provides strong evidence for machine learning effectiveness, operational deployment would require addressing implementation challenges, including data quality assurance, institutional capacity building, and integration with existing health information systems.

Fifth, the study used confirmed case counts rather than incidence rates (cases per population), which may not fully account for population changes over time or differences between LGAs. Future work should incorporate population denominators when available.

4.2 CONCLUSION

This study has demonstrated that machine learning models, particularly Random Forest, can effectively capture the nonlinear, time-dependent relationships between climatic factors and malaria transmission in Kebbi State, significantly outperforming conventional statistical approaches. The Random Forest regression model achieved near-perfect predictive accuracy ($R^2 = 0.9967$) with minimal error, while the classification model demonstrated strong generalization capability (89.52% test accuracy) for identifying high-risk periods.

The findings confirm that rainfall and humidity, particularly their lagged effects, are the dominant climatic predictors of malaria incidence. The biological delay between environmental changes and increased transmission highlights the critical importance of incorporating temporal dynamics into predictive modeling frameworks.

This study provides a robust foundation for developing operational malaria early warning systems in Kebbi

State and similar settings. The integration of climatic and epidemiological data with machine learning analytics offers a transformative opportunity for proactive, data-driven malaria control strategies.

4.3 FUTURE WORK

Several directions for future research and development emerge from this study:

1. **Incorporation of Additional Variables:** Future models should integrate socioeconomic factors (population density, housing quality, income levels), environmental variables (vegetation indices, land use, water bodies), and intervention data (ITN distribution, IRS coverage, treatment seeking behavior) to enhance predictive accuracy and explanatory power.
- 2.
3. **Spatial Modeling:** Extending the analysis to incorporate spatial dependencies between LGAs would enable identification of transmission hotspots and improved understanding of disease spread patterns.
- 4.
5. **Real-Time Prediction System:** Development of a web-based or mobile application that automatically

ingests meteorological data and generates malaria risk alerts for health officials.

- i. **Transfer Learning:** Investigation of model transferability to other Nigerian states or West African countries with similar climatic profiles.
- ii. **Longer-Term Forecasting:** Exploration of models capable of predicting malaria incidence at longer lead times (3-6 months) to enable more proactive intervention planning.
- iii. **Integration with Seasonal Climate Forecasts:** Using seasonal climate forecasts rather than observed data would extend the prediction horizon and enable pre-season preparedness.
- iv. **Ensemble Model Approaches:** Combining RF, ANN, and other models through ensemble methods to potentially achieve even better performance.
- v. **Explainable AI:** Development of model interpretation tools to help public health officials understand prediction drivers and build trust in AI-based recommendations.

REFERENCES

- Adamu, Y. A., & Singh, J. (2021). Malaria prediction model using machine learning algorithms. *Turkish Journal of Computer and Mathematics Education*, 12(10), 7488-7496.
- Amadi, M., & Erandi, K. H. (2024). Assessing the relationship between malaria incidence levels and meteorological factors using cluster-integrated regression. *BMC Infectious Diseases*, 24(1), 664. <https://doi.org/10.1186/s12879-024-09567-8>
- Bayoh, M. N., & Lindsay, S. W. (2003). Temperature-related duration of aquatic stages of the Afrotropical malaria vector mosquito *Anopheles gambiae* in the laboratory. *Medical and Veterinary Entomology*, 17(2), 197-204. <https://doi.org/10.1046/j.1365-2915.2003.00424.x>
- Caminade, C., McIntyre, K. M., & Jones, A. E. (2019). Climate variability and its impact on malaria incidence dynamics in Africa. *Environmental Research Letters*, 14(2), 020301. <https://doi.org/10.1088/1748-9326/aaf981>
- Lee, Y. W., Choi, J. W., & Shin, E. H. (2021). Machine learning model for predicting malaria using clinical information. *Computers in Biology and Medicine*, 129, 104161. <https://doi.org/10.1016/j.compbiomed.2020.104161>
- Nanvyat, N., Mulambalah, C. S., Barshep, Y., Ajiji, J. A., Dakul, D. A., & Tsingalia, H. M. (2018). Malaria incidence trends and its lagged association with climatic factors in the highlands of Plateau State, Nigeria. *Journal of Tropical Parasitology*, 8(1), 18-23. <https://doi.org/10.4103/2229-5070.239932>
- Nyawanda, B. O., Beloconi, A., Khagayi, S., Bigogo, G., Obor, D., Otieno, N. A., ... & Vounatsou, P. (2023). The relative effect of climate variability on malaria incidence after scale-up of interventions in western Kenya: A time-series analysis of monthly incidence data from 2008 to 2019. *Parasite Epidemiology and Control*, 21, e00297. <https://doi.org/10.1016/j.parepi.2023.e00297>
- Ogunsakin, R. E., Babalola, B. T., Olusola, J. A., Joshua, A. O., & Okpeku, M. (2024). GIS-based spatiotemporal mapping of malaria prevalence and exploration of environmental inequalities. *Parasitology Research*, 123(7), 262. <https://doi.org/10.1007/s00436-024-08268-6>
- Oheneba-Dornyo, T. V., Amuzu, S., Maccagnan, A., & Taylor, T. J. (2022). Estimating the impact of temperature and rainfall on malaria incidence in Ghana from 2012 to 2017. *Environmental Modeling & Assessment*, 27(3), 473-489. <https://doi.org/10.1007/s10666-021-09812-7>
- Okundalaye, O., Ozdemir, N., Rotimi, B., & Akanbi, F. (2025). Climate-based predictive modeling of malaria incidence using statistical and machine learning approaches. *Transactions on Computational Modelling and Intelligent Systems*, 1, 10014.
- Paaijmans, K. P., Read, A. F., & Thomas, M. B. (2009). Understanding the link between malaria risk and climate. *Proceedings of the National Academy of Sciences*, 106(33), 13844-13849. <https://doi.org/10.1073/pnas.0903423106>
- Padhi, B. K., Gaidhane, A. M., Satapathy, P., Bushi, G., Ballal, S., Bansal, P., ... & Shabil, M. (2024). Assessing the impact of ecological, climatic, and socioeconomic factors on age-specific malaria



DOI: <https://doi.org/10.57233/ijsgs.v12i3.1182>

ISSNp: 2488-9229; ISSNc: 3027-1118

IJSGS FUGUSAU VOL. 12(3)

WEBSITE: <https://fugus-ijsgs.com.ng>

incidence in India: A mixed-model approach using the Global Burden of Disease Study (2010-2019). *Malaria Journal*, 23(1), 332. [https://doi.org/10.1186/s12936-](https://doi.org/10.1186/s12936-024-05154-3)

[024-05154-3](https://doi.org/10.1186/s12936-024-05154-3)

Uzun Ozsahin, D., Duwa, B. B., Ozsahin, I., & Uzun, B. (2024). Quantitative forecasting of malaria parasite

using machine learning models: MLR, ANN, ANFIS and Random Forest. *Diagnostics*, 14(4), 385. <https://doi.org/10.3390/diagnostics14040385>

World Health Organization. (2024). *World Malaria Report 2024*. Geneva: WHO.