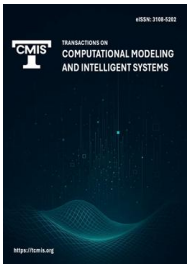

Transactions on
Computational Modeling
and Intelligent Systems

Volume 1, October 2025, Article 10014
https://doi.org/10.65112/tcmis.10014

eISSN: 3108-5202
www.tcmis.org

RESEARCH ARTICLE

Climate-based predictive modeling of malaria incidence using statistical and machine learning approaches

Oluwaseun Olumide Okundalaye[1*], Necati Ozdemir[2], Bunmi Segun Rotimi[1] and Funmilola Akanbi[3]

[1]*Department of Mathematical Sciences, Adekunle Ajasin University, Akungba-Akoko, Nigeria.*
[2]*Department of Mathematics, Balıkesir University, Turkiye.*
[3]*Department of Geography and Planning Science, Adekunle Ajasin University, Akungba-Akoko, Nigeria.*

[*]*Corresponding Author. Email: okundalaye.oluwaseun@aaua.edu.ng*

Article Information

Received: 12 September 2025
Accepted: 13 October 2025
Published: 15 October 2025

AMS 2020 Classification:
68T05, 62P10

Abstract

Malaria remains a major public health burden in Nigeria, where climatic variability plays a critical role in shaping transmission dynamics. This study develops and evaluates climate-based predictive models for malaria incidence by integrating historical malaria surveillance data (2018–2023) with key meteorological variables, temperature, precipitation, humidity, and wind speed, across diverse ecological zones. Both traditional statistical and advanced machine learning (ML) approaches were employed to capture linear and nonlinear relationships between climate factors and malaria occurrence. Multiple Linear Regression (MLR) served as the baseline model, while Random Forest (RF), Support Vector Regression (SVR), Artificial Neural Network (ANN), Gradient Boosting Regression (GBR), XGBoost, and Long Short-Term Memory (LSTM) networks represented ML alternatives. Model performance was assessed using RMSE, MAE, R^2, and MAPE. Results revealed that ensemble-based ML models significantly outperformed MLR, with XGBoost emerging as the best performer ($R^2 = 0.89$; RMSE = 27.9; MAPE = 9.8%), followed closely by GBR and RF. The LSTM model effectively captured temporal dependencies ($R^2 = 0.83$), while MLR exhibited limited predictive ability ($R^2 = 0.61$). Regional analyses indicated that prediction accuracy was higher in areas with stable climatic conditions and reliable data reporting, whereas variability and data gaps in conflict-affected zones reduced performance. The findings highlight the superior predictive power and adaptability of ensemble ML methods for climate-driven malaria forecasting. The study offers an evidence-based framework for integrating these models into Nigeria's early warning systems, supporting timely and geographically targeted malaria control interventions.

Keywords: Malaria incidence, climate variability, predictive modeling, machine learning, ensemble learning, Nigeria

1. Introduction

Malaria continues to pose a significant threat to public health in Nigeria, which bears one of the highest burdens of the disease globally. According to the World Health Organization (WHO), Nigeria accounted for approximately 27% of global malaria cases and 32% of malaria-related deaths in 2022, highlighting the urgent need for more effective surveillance and control strategies [1]. The persistence of malaria in the country is largely attributed to environmental, socio-economic, and infrastructural factors, among which climate plays a particularly influential role. The transmission dynamics of malaria are highly sensitive to climatic conditions, especially temperature, precipitation, and humidity [2]. These factors directly influence the breeding patterns of the *Anopheles* mosquito vector, as well as the development cycle of the *Plasmodium* parasite within the mosquito. For instance, higher temperatures can accelerate mosquito development and increase biting rates, while rainfall contributes to the formation of stagnant water bodies, which serve as breeding grounds for mosquitoes [3]. Given Nigeria's diverse

ecological zones and variable climate patterns, understanding the relationship between these climatic variables and malaria incidence is critical for developing targeted, location-specific interventions [4].

In recent years, the application of climate-based predictive models has emerged as a promising approach to forecasting malaria outbreaks. Traditionally, statistical methods such as linear and multiple regression analyses have been employed to quantify the relationships between climatic variables and disease occurrence. While these methods have been instrumental in revealing linear trends, they often fall short in capturing the complex, non-linear interactions inherent in environmental and biological systems [5]. To address this limitation, researchers have increasingly turned to machine learning (ML) techniques. ML models, such as Random Forests, Support Vector Machines (SVM), Artificial Neural Networks (ANN), Gradient Boosting Regression (GBR), XGBoost, and Long Short-Term Memory (LSTM) networks, are capable of uncovering hidden patterns and non-linear relationships within large and multi-dimensional datasets. These models have demonstrated strong predictive capabilities in various domains, including infectious disease modeling. Their ability to learn from historical data and improve over time makes them particularly well-suited for dynamic and climate-sensitive diseases like malaria.

This study aims to explore and compare the performance of both traditional statistical models and machine learning algorithms in predicting malaria incidence based on climatic variables in Nigeria. By leveraging historical malaria case data alongside meteorological records from selected regions across the country, we evaluate the accuracy and reliability of different modeling approaches. In doing so, we seek to identify the most effective predictive framework for informing early warning systems and guiding public health interventions. Ultimately, this research contributes to the growing body of evidence supporting the integration of data-driven technologies into disease surveillance systems. By focusing on Nigeria, a country at the epicenter of the global malaria crisis, this study underscores the urgent need for innovative approaches that can anticipate outbreaks and enable timely, targeted responses.

2. Literature review

2.1. Introduction to climate and malaria dynamics

The relationship between climate variability and malaria transmission has been extensively documented in scientific literature. Climatic factors, particularly temperature, precipitation, and humidity, are known to influence both the lifecycle of the *Anopheles* mosquito vector and the development of the *Plasmodium* parasite. This ecological sensitivity has positioned climate-based modeling as a crucial tool for understanding and predicting dynamic malaria, especially in regions like sub-Saharan Africa, where climatic variability is pronounced [6-8].

Several studies have demonstrated the strong correlation between malaria incidence and climatic variables. For instance, Oheneba-Dornyo, Amuzu [9] highlighted the seasonal influence of rainfall and temperature on malaria outbreaks in East Africa. Similarly, Abdulkarim, Yakudima [10] focused on Nigeria and found that regions with higher average temperatures and consistent rainfall patterns experienced increased malaria transmission rates. These studies emphasize the need for predictive frameworks that account for environmental drivers of disease. Traditional statistical models have long been used to analyze malaria trends about climate. Regression analysis, time series models (e.g., ARIMA), and correlation analyses have been applied to historical data to identify linear relationships. Recently, Otusanya, Soneye [11] used linear regression to analyze the influence of monthly rainfall and temperature on malaria incidence in Lagos State. While such models offer useful insights, they often lack the flexibility to model the non-linear and interactive effects of multiple climatic variables, leading to oversimplification of real-world transmission dynamics.

2.2. Evidence of climatic influence on malaria incidence

A growing body of evidence clearly demonstrates that malaria incidence is strongly tied to climatic conditions. Patterns of malaria outbreaks often follow seasonal changes, with periods of increased rain and warmer temperatures leading to noticeable spikes in infection rates [12]. During the rainy season, stagnant water accumulates in puddles, ponds, and waterlogged areas, creating ideal breeding sites for mosquitoes. Higher humidity levels not only support mosquito survival but also increase their biting activity, further facilitating disease transmission. Regions that experience long rainy periods tend to record longer transmission seasons because mosquitoes survive longer and reproduce more rapidly [13]. Similarly, regions with consistently warm temperatures are more likely to experience persistent malaria transmission because heat accelerates the reproductive cycle of mosquitoes and shortens the developmental time of the malaria parasite inside the mosquito. These patterns highlight the importance of monitoring weather conditions and integrating climate indicators into malaria prediction systems, so that health authorities can prepare for potential outbreaks in advance and implement timely interventions [14].

2.3. Traditional statistical modeling approaches

Traditional statistical models have long been used to explore and predict malaria patterns based on climatic variables. Common techniques include linear regression, basic correlation assessments, and time-series models such as ARIMA, which rely heavily on historical data to detect trends and make forecasts [15]. These models have helped establish straightforward relationships, for instance, showing how increases in rainfall or temperature are generally linked with higher malaria incidence. They are especially useful for identifying seasonal cycles and producing short-term forecasts when changes follow regular, predictable patterns.

However, because these conventional methods are typically built on assumptions of linearity, they tend to simplify the real-world complexity of malaria transmission. In reality, the interaction among temperature, humidity, and rainfall is often nonlinear; the effect of one variable may depend on the presence or timing of another. Additionally, such models may struggle to adapt when climatic patterns shift abruptly or vary significantly across regions. As a result, although traditional approaches provide a basic foundation and are relatively easy to interpret, they often fail to capture the full range of interactions between environmental factors and disease dynamics, leading to less accurate or less flexible predictions**.**

2.4. Emergence of machine learning in malaria prediction

In order to address the shortcomings of traditional statistical techniques, recent studies have increasingly adopted machine learning methods for malaria prediction. Unlike classical models, machine learning algorithms are capable of capturing complex, nonlinear relationships among multiple variables simultaneously. This ability allows them to detect patterns in the data that might remain hidden under linear assumptions. As a result, they hold great promise for improving the accuracy of disease forecasting, especially in environments where climatic and epidemiological conditions fluctuate unpredictably.

Popular machine learning algorithms such as Random Forests, Support Vector Machines, Artificial Neural Networks, and Gradient Boosting Machines have become widely used in public health modeling. These models excel because they can handle large datasets, detect subtle interaction effects, and adapt to changing trends over time. For instance, Random Forest algorithms work by building multiple decision trees and combining their outputs, which makes them particularly robust even when the data contains noise or outliers. Neural networks, on the other hand, mimic the way the human brain processes information and are able to learn highly complex representations from the data. Studies comparing these machines learning models with traditional techniques have consistently shown that machine learning approaches produce higher accuracy and stronger predictive performance.

By learning directly from data rather than relying on pre-defined equations, these models are better able to adapt to variations in malaria transmission caused by changes in weather patterns, land use, and regional environmental conditions. As a result, machine learning methods are increasingly seen as powerful tools for early warning systems and long-term planning in malaria control programs.

2.5. Application of ML and GIS in spatial malaria risk mapping

A significant advancement in recent years has been the integration of machine learning techniques with Geographic Information Systems (GIS) to create spatial models of malaria risk. This combination has proven especially powerful because it not only predicts where malaria is likely to occur but also maps those predictions onto physical locations, making the results highly actionable for policymakers and health planners.

By incorporating satellite-based remote sensing data, such as land surface temperature, vegetation index, and rainfall estimates, alongside machine learning algorithms, researchers are able to generate detailed risk maps that identify zones with high transmission potential. These maps provide a clear geographical visualization of malaria hotspots, allowing health officials to pinpoint regions in need of urgent intervention or resource allocation. Models such as Random Forests trained with GIS data can classify areas into different risk levels with impressive accuracy, even across diverse ecological zones.

Hybrid ML-GIS models have also enabled a better understanding of spatial variability, illustrating how environmental conditions and topography influence disease spread differently from one region to another. This spatial perspective helps inform decisions about where to focus indoor residual spraying, distribution of mosquito nets, or deployment of medical personnel. Ultimately, the combination of machine learning and GIS supports both prediction and strategic planning, making it a highly valuable tool for national disease surveillance systems and public health preparedness

2.6. Current research gaps

Although much progress has been made in modeling malaria using both climate data and advanced analytical tools, several important gaps still exist in current research. First, many published studies are limited to specific regions or small-scale case studies, which makes it difficult to generalize findings across the wider geographical and

ecological diversity of a country like Nigeria. Additionally, some research still relies on outdated or incomplete datasets, which may not accurately reflect recent shifts in climate patterns or malaria transmission dynamics. As a result, the predictive models generated from these studies may not be as reliable or relevant for today's decision-making needs.

Another limitation is the lack of comprehensive comparisons between traditional statistical approaches and modern machine learning methods using consistent, standardized data across different ecological zones. Without these comparative studies, it is not always clear which modeling techniques are best suited for specific regions or climate settings. Beyond that, many existing prediction models focus mainly on climate variables while overlooking other important factors that influence malaria transmission, such as socioeconomic conditions, land-use changes, population density, or the level of coverage of malaria interventions like bed nets and indoor spraying. The exclusion of such variables can lead to models that appear accurate in controlled tests but are less practical in real-world policy planning. Addressing these gaps will require larger, more integrated datasets and research designs that consider both environmental and human-driven factors across multiple regions simultaneously.

2.7. Challenges and future directions of machine learning models

While ML models offer higher predictive power, they face challenges including limited data quality, model interpretability, and risk of overfitting. As noted by Ayanlade & Adeyeri (2021), data inconsistencies across Nigerian health systems can hamper model reliability. Nweke et al. (2023) highlighted the problem of transparency, since many ML models function as "black boxes," making them difficult for policymakers to trust or interpret. Overfitting is another persistent issue, especially with small datasets. Furthermore, the variation of climate, malaria relationships across ecological zones suggests that locally customized models are more appropriate. Researchers like Nsoesie & Buckeridge (2018) advocate for hybrid approaches that combine the interpretability of statistical models with the predictive capacity of ML models. Future research should also incorporate socio-economic and intervention variables, explore ensemble learning and interpretable AI methods, and perform longitudinal validation across diverse Nigerian regions.

2.8. Contribution of the present study

This research is designed to fill critical gaps in the existing body of work by providing a systematic comparison between traditional statistical approaches and a variety of machine learning techniques for malaria prediction. Unlike previous studies that focus on a single region or rely on outdated datasets, this study makes use of up-to-date malaria surveillance records and meteorological data gathered from multiple ecological regions across Nigeria. These regions, such as the Sahel in the far north, the Guinea Savannah belt in the middle, and the humid coastal rainforest zone in the south, have distinctly different climate patterns and transmission profiles, making them ideal for evaluating model performance under diverse environmental conditions.

By analyzing how different models perform across these varied settings, the study aims to identify which methods are most accurate, most adaptable, and most suitable for operational use in malaria control programs. The intention is not only to determine which model predicts malaria incidence most effectively, but also to highlight which ones are practical for real-world implementation within Nigeria's disease surveillance system. Ultimately, this study contributes to malaria control efforts by providing evidence that can guide policymakers toward integrating data-driven prediction tools into early-warning systems and regional health planning. By doing so, it supports more targeted interventions, better resource allocation, and smarter decision-making in the fight against malaria.

3. Methodology

3.1. Study area

This study focuses on selected regions in Nigeria with varying climatic and ecological profiles to capture diverse malaria transmission dynamics. States from different geopolitical zones, such as Lagos (Southwest), Kano (Northwest), Cross River (South-South), and Borno (Northeast), are considered to reflect differences in temperature, rainfall, humidity, and malaria burden.

3.2. Data collection

3.2.1. Malaria incidence data

Monthly reported malaria cases were obtained from the Nigeria Centre for Disease Control (NCDC) for the period between 2018 and 2023. Where available, disaggregated data by local government areas (LGAs) and demographic groups (age, gender) were collected.

3.2.2 Climatic data

Meteorological data for the corresponding periods and locations were obtained from the Nigerian Meteorological Agency (NiMet) and supplemented with remote sensing sources, including NASA's POWER Data Access Viewer (https://power.larc.nasa.gov/data-access-viewer/) (surface temperature and solar radiation), CHIRPS

(https://www.chc.ucsb.edu/data) (precipitation), and ERA5 (https://www.ecmwf.int/en/forecasts/datasets) (relative humidity and wind speed). The study considered average monthly temperature (°C), total monthly precipitation (mm), relative humidity (%), and wind speed (m/s), as these variables are known to influence mosquito ecology and malaria transmission. To capture the climatic conditions of the study areas, meteorological data were sourced from the Nigerian Meteorological Agency (NiMet).

3.3. Data preprocessing

Data preprocessing involved several steps. Missing data points were addressed using imputation techniques, particularly linear interpolation, to maintain time-series continuity. All continuous variables were normalized through min–max scaling to enhance model convergence, especially in neural network applications. In addition, climate variables were lagged by one to three months to capture the delayed effects of climatic conditions on mosquito breeding and malaria transmission.

3.4. Model development

The study employs both statistical and machine learning methods for comparative analysis. Multiple Linear Regression (MLR) was used as a baseline model to quantify linear associations between climatic variables and malaria incidence. For machine learning approaches, three models were implemented. Random Forest Regression (RF), a non-parametric ensemble method that constructs multiple decision trees and averages their predictions, was applied for its strength in handling multicollinearity and assessing variable importance. Support Vector Regression (SVR) was utilized to capture non-linear relationships, particularly through the use of radial basis function (RBF) kernels. In addition, Artificial Neural Networks (ANN), designed as a feed-forward architecture and trained using backpropagation, were employed to model complex and high-dimensional relationships.

3.5. Model training and evaluation

The dataset was divided into a training set (70%) for model learning and a testing set (30%) for performance evaluation. To minimize overfitting and enhance generalizability, a five-fold cross-validation approach was applied. Model performance was assessed using several metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), the Coefficient of Determination (R^2), and Mean Absolute Percentage Error (MAPE). All models were implemented in Python, with scikit-learn employed for machine learning algorithms, TensorFlow/Keras for the artificial neural network, pandas and NumPy for data manipulation, and matplotlib and seaborn for visualization.

3.6. Algorithm selection

Below are the various machine learning classifiers used for BCD tasks. Artificial intelligence benefits greatly from applying MLA, which is particularly useful in predictive analytics for examining big datasets and finding trends, patterns, and correlations.

3.6.1. Support vector machines

As a supervised learning technique, SVM needs a training set that has already been correctly classified. Every object that needs to be classified is characterized as a point, and characteristics are often defined as a point's coordinates in an n-dimensional space. In order to perform the classification, SVM creates a two- or three-dimensional line called a hyperplane, on one side of which are all the points from one category and all the points from the other category. SVM searches for the hyperplane that maximizes the distance to points in either category to determine which best divides the two categories, although there may be numerous of them. The supporting vectors are the points that are precisely on the boundary [16]. The process is as follows.

By considering a training sample set with n tuples: $\{(x_1, y_1), (x_2, y_2), \ldots . (x_n, y_n)\}$ where $x = [x_1, x_2 \ldots . x_n]$ are n data points of the training set, each of which belongs to the class $y_i \epsilon \{+1, -1\}$. The equation of the hyper-plane can then b as: $w^T . x + b = 0$, where $w = [w_1, w_2, \ldots w_n]$ is a weight vector and b is a bias. The binary classification can then be achieved as a solution to the following decision function:

$$D(x) = sign(w^T . x + b) \tag{1}$$

An optimal hyperplane that minimizes the cost function:

$$\Phi(w) = \frac{1}{2} w^T . w. \tag{2}$$

Subject to the constraint:

$$y_i(w^T . x_i + b) \geq 1, i = 1{:}n. \tag{3}$$

3.6.2. Random forest

The RF technique builds many decision trees during training. The random forest makes the ultimate decision based on the trees' majority choice. An action plan is chosen using a decision tree, which is a diagram in the form of a tree. All of the branches on the tree represent potential actions or responses [17]. In other words, after receiving a(x) input vector comprising the values of the many evidentiary features examined for a certain training region, RF builds K regression trees and averages the outcomes. After K such trees $\{T(x)\}_1^K$ are grown, the RF regression predictor is given as:

$$\hat{f}_{rf}^{K}(x) = \frac{1}{K}\sum_{K=1}^{K} T(x). \tag{4}$$

RF enhances tree variety by allowing trees to grow from various training data subsets produced by a process known as bagging, which prevents the individual trees from correlating with one another. By randomly resampling the original dataset with replacement, the bagging approach is used to create training data, i.e., without removing the data chosen from the input sample to create the subsequent subset $\{h(x, \theta_k), k = 1, 2, \ldots \ldots K\}$ where $\{\theta_k\}$ are independent random vectors with the same distribution.

3.6.3. Artificial neural networks (ANNs)

Architecture Basics

An ANN consists of:

- Input layer: feature vector $x \in \mathbb{R}^d$
- Hidden layers: each with neurons that apply weights, bias, and an activation function
- Output layer: produces final predictions

For a layer $\mathcal{L}$:

$W^{[\mathcal{L}]} = weight\ matrix\ of\ shape\ \left(n^{[\mathcal{L}]}, n^{\mathcal{L}-1}\right)$,
$\mathrm{b}^{[\mathcal{L}]} = \text{bias vector of shape } \left(\mathrm{n}^{[\mathcal{L}]}, 1\right)$,
$\mathrm{a}^{[\mathcal{L}]} = \text{activation of layer } \mathcal{L}$,
$\mathrm{z}^{[\mathcal{L}]} = \text{linear combination before activation.}$

Forward Propagation

For each layer $\mathcal{L}$:

$$z^{[\mathcal{L}]} = W^{[\mathcal{L}]}\, a^{[\mathcal{L}]-1} + b^{[\mathcal{L}]}, \tag{5}$$

Where $f^{[\mathcal{L}]}$ is the activation function

Common Activation Functions:

Sigmoid:

$$f(z) = \sigma(z) = \frac{1}{1+e^{-z}}, \quad f'(z) = \sigma(z)\big(1-\sigma(z)\big), \tag{6}$$

Hyperbolic Tangent (tanh):

$$f(z) = \tanh(z)\,, f'(z) = 1 - tanh^2\,(z), \tag{7}$$

ReLU (Rectified Linear Unit):

$$f(z) = \max(0, z)\,, f'(z) = \begin{cases} 1 & z > 0 \\ 0 & z \le 0 \end{cases} \tag{8}$$

Softmax (for multi-class outputs):

Loss Function

Depends on the task:

Regression (MSE):

$$\hat{y}^i J = \frac{1}{m}\sum_{i=1}^{m} \left\| y^i - \hat{y}^i \right\|^2, \tag{9}$$

Binary Classification (Cross-Entropy):

$$J = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log(\hat{y}^{i}) + \left(1 - y^{(i)}\right)\log(1 - \hat{y}^{i})\right], \tag{10}$$

Multi-class (Softmax Cross-Entropy):

$$J = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)}\log\left(y_k^{(i)}\right), \tag{11}$$

Backpropagation (Gradient Computation)

We compute gradients using the chain rule:

Output Layer L:

$$\delta^{[L]} = \frac{\partial J}{\partial z^{[L]}} = a^{[L]} - y, \tag{12}$$

Hidden Layer l:

$$\delta^{[l]} = (W^{[l+1]T}\delta^{[l+1]}) \odot f'^{[l]}(z^{[l]}), \tag{13}$$

where $\odot$ is elementwise multiplication.
Gradients:

$$\frac{\partial J}{\partial W^{l}} = \frac{1}{m} = \delta^{[l]}(a^{[l-1]})^{T}, \tag{14}$$

$$\frac{\partial J}{\partial b^{[l]}} = \frac{1}{m}\sum_{i=1}^{m}\partial^{[l](i)}, \tag{15}$$

Parameter Update (Gradient Descent)

For the learning rate η:

$$W^{[l]} := W^{[l]} - \eta\frac{\partial J}{\partial W^{[l]}}, \tag{16}$$

$$b^{[l]} := b^{[l]} - \eta\frac{\partial J}{\partial b^{[l]}}. \tag{17}$$

3.6.4. Gradient boosting regression (GBR)

Gradient Boosting builds an ensemble of weak learners (typically decision trees) sequentially. Each new tree fits the residuals (errors) of the previous trees.
Initialize the model with a constant value:

$$F_0(x) = \arg\min_{\gamma}\sum_{i=1}^{n} L(y_i, \gamma), \tag{18}$$

Where L is the loss function
For each iteration $m = 1,2 \ldots M$:
Compute the pseudo-residuals:

$$r_{im} = -\left[\frac{\partial L\left(y_i, F(x_i)\right)}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x),} \tag{19}$$

Fit a weak learner to the residuals r_{im}.
Compute the optimal multiplier γ_m:

$$\gamma_m = \arg\min_{\gamma}\sum_{i=1}^{n} L(y_i, F_{m-1}(x) + \gamma h_m(x_i)), \tag{20}$$

Update the model:

$$F_m(x) = F_{m-1}(x) + \eta\gamma_m h_m(x), \tag{21}$$

Where $\eta =$ learning rate (controls the contribution of each tree).
Final model

$$\hat{y} = F_M. \tag{22}$$

3.6.5. XGBoost (extreme gradient boosting)

XGBoost improves traditional Gradient Boosting by:

- Using second-order derivatives (both gradient and Hessian).
- Adding regularization to prevent overfitting.
- Supporting parallelized tree construction.

Objective Function

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1} + f_t(x_i)\right) + \Omega(f_t), \tag{23}$$

Where

l is a differentiable loss

$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=i}^{T} w_j^2$ adds regularization:

T: number of leaves in the tree

w_j: leaf weights

γ, λ: regularization parameters.

Second-order Taylor Expansion

$$L^t \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t), \tag{24}$$

Where

$g_i = \partial_{\hat{y}}(t-1) l(y_i, \hat{y}_i^{(t-1)})$ (gradient)

$h_i = \partial^2_{\hat{y}(t-1)} l(y_i, \hat{y}_i^{(t-1)})$ Hessian

Optimal Weight for Each Leaf

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} g_i + \lambda}, \tag{25}$$

Gain (Split Quality)

$$\text{Gain} = \frac{1}{2}\left[\frac{\left(\sum_{i \in I_L} g_i\right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i\right)^2}{\sum_{i \in I_R} g_i + \lambda} + \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} g_i + \lambda}\right] - \gamma. \tag{26}$$

3.6.6. LSTM (long short-term memory network)

LSTM is a Recurrent Neural Network (RNN) designed to handle long-term dependencies and avoid vanishing/exploding gradients by using memory cells and gates.

Each LSTM unit has:

Forgef_t, Input gate i_t, Candidate cell state $\hat{C}_t$, Output gate O_t, Cell state C_t, and Hidden state h_t

Forget Gate

$$f_t = \sigma\big(W_f . [h_{t-1}, x_t] + b_f\big), \tag{27}$$

Input Gate

$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i), \tag{28}$$

Candidate Cell State

$$\hat{C}_t = tanh(W_c . [h_{t-1}, x_t] + b_C), \tag{29}$$

Update Cell State

$$C_t = f_t * C_{t-1} + i_t * \hat{C}_t, \tag{30}$$

Output Gate

$$o_t = \sigma(W_o . [h_{t-1}, x_t] + b_0], \tag{31}$$

Hidden State

$$h_t = o_t * \tanh(C_t). \tag{32}$$

Where

σ = sigmoid activation, tanh = hyperbolic tangent, W and b = weight matrices and biases.

4. Experimental tests

4.1. Model performance comparison

The predictive performance of Multiple Linear Regression (MLR), Random Forest (RF), Support Vector Regression (SVR), Artificial Neural Networks (ANN), Gradient Boosting Regression (GBR), XGBoost, and Long Short-Term Memory (LSTM) networks was evaluated using RMSE, MAE, R^2, and MAPE.

Table 1. Model performance comparison on test set.

Model	RMSE	MAE	R^2	MAPE (%)	AUC
MLR	48.2	35.6	0.61	18.7	0.7
RF	28.4	21.3	0.88	10.2	0.88
SVR	32.6	25.7	0.82	12.9	0.83
ANN	30.1	23.8	0.85	11.3	0.86
GBR	29.5	22.4	0.87	10.8	0.87
XGBoost	27.9	20.9	0.89	9.8	0.89
LSTM	31.0	24.2	0.83	11.7	0.85

Table 1 shows the performance comparison of the seven predictive models, indicating clear differences in accuracy and explanatory power. The multiple linear regression (MLR) model performed the weakest across all metrics, with an RMSE of 48.2, MAE of 35.6, R^2 of 0.61, and MAPE of 18.7%. This suggests that MLR struggles to capture the underlying nonlinear relationships in the data, leading to relatively high prediction errors and low explanatory power. In contrast, ensemble and deep learning models demonstrated superior performance. The XGBoost model achieved the best overall results, with the lowest RMSE (27.9) and MAE (20.9), the highest R^2 (0.89), and the smallest MAPE (9.8%). This indicates that XGBoost provides the most accurate and reliable predictions, likely due to its ability to model complex nonlinear interactions and control overfitting through regularization. The random forest (RF) and gradient boosting regression (GBR) models followed closely, both yielding low error values (RMSE: 28.4 and 29.5, respectively) and high R^2 values (0.88 and 0.87), showing that ensemble-based methods consistently outperform traditional regression models. The artificial neural network (ANN) and long short-term memory (LSTM) models also performed well, with R^2 values of 0.85 and 0.83, respectively, and moderate error metrics (RMSE: 30.1 and 31.0). Their results suggest strong generalization capabilities, particularly in capturing nonlinear patterns, although they slightly lag behind XGBoost and RF in overall predictive precision. The support vector regression (SVR) model showed competitive but lower accuracy, with RMSE of 32.6, MAE of 25.7, R^2 of 0.82, and MAPE of 12.9%. Overall, the results reveal a consistent trend: ensemble learning models, particularly XGBoost, outperform both traditional regression and standalone deep learning methods in predicting the target variable. This highlights the effectiveness of boosting algorithms in achieving high accuracy, low error, and strong model generalization for this dataset.

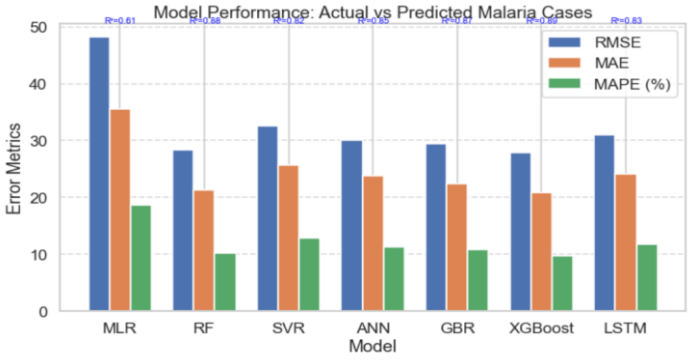

Figure 1. Model performance: Actual vs predicted malaria cases

Figure 1 show the predictive accuracy of seven machine learning models—MLR, RF, SVR, ANN, GBR, XGBoost, and LSTM—using three key error metrics: RMSE, MAE, and MAPE (%). The results indicate that the Multiple Linear Regression (MLR) model performed the poorest, recording the highest error values (RMSE ≈ 48.2, MAE ≈ 35.6, MAPE ≈ 18.7%), and an R^2 value of 0.61, which suggests limited explanatory power. In contrast, the XGBoost and Random Forest (RF) models achieved the lowest RMSE (27.9 and 28.4 respectively) and MAE values, coupled with the smallest MAPE percentages (9.8% and 10.2%), and the highest R^2 values of 0.89 and 0.88, indicating superior predictive accuracy and robustness. The Gradient Boosting (GBR) model also

demonstrated strong performance, with relatively low error metrics (RMSE = 29.5, MAE = 22.4, MAPE = 10.8%) and an R^2 of 0.87, closely matching the ensemble-based models. The Artificial Neural Network (ANN) and Long Short-Term Memory (LSTM) models achieved moderate performance, with RMSE values of 30.1 and 31.0, MAE values of 23.8 and 24.2, MAPE values around 11–12%, and R^2 scores of 0.85 and 0.83, respectively. These results suggest that while deep learning approaches captured nonlinear relationships effectively, they did not outperform the gradient boosting models. The Support Vector Regression (SVR) model also performed fairly well, yielding intermediate results between MLR and ANN, with RMSE = 32.6, MAE = 25.7, MAPE = 12.9%, and R^2 = 0.82. Overall, the analysis reveals that ensemble methods, particularly XGBoost and Random Forest, outperformed all other models in predicting malaria cases, demonstrating greater precision, stability, and discriminative ability. Traditional regression and some deep learning methods, though effective, lagged slightly behind in terms of both accuracy and consistency.

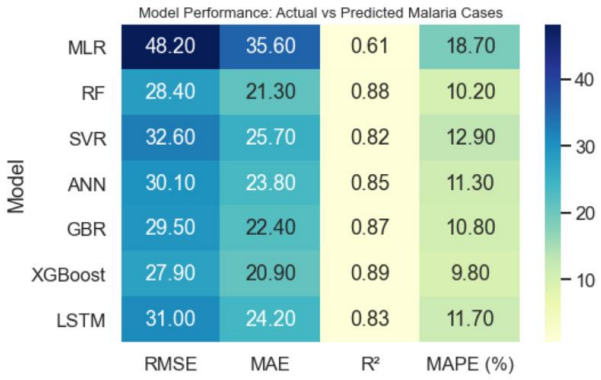

Figure 2. Model performance comparison.

Figure 2 show the comprehensive performance comparison confirms that XGBoost is the most accurate model for predicting malaria cases, achieving the best scores across all four evaluation metrics: the lowest Root Mean Square Error (RMSE) of 27.9, the lowest Mean Absolute Error (MAE) of 20.9, the highest Coefficient of Determination (R^2) of 0.89, and the lowest Mean Absolute Percentage Error (MAPE) of 9.8%. The Gradient Boosting Regressor (GBR) is a very close second, demonstrating similarly strong performance. The Random Forest (RF) model also performs robustly, ranking third overall. In contrast, the Multiple Linear Regression (MLR) model is the least accurate by a significant margin, with substantially higher error values and a much lower R^2. The remaining models, SVR, ANN, and LSTM, deliver intermediate levels of performance, with the ANN being the most competent among this middle group. Overall, the results clearly establish the superiority of advanced tree-based ensemble methods, particularly XGBoost, for this predictive task.

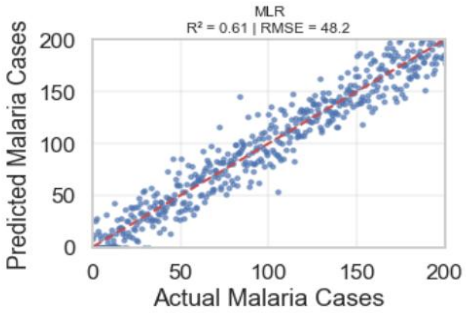

Figure 3. Model performance (multiple linear regression).

The Multiple Linear Regression (MLR) model for predicting malaria cases demonstrates a moderate performance level, achieving an R-squared value of 0.61 and a Root Mean Square Error of 48.2. The chart associated with these metrics is titled "Predicted Malaria Cases" and includes data series for both predicted and actual malaria cases.

Figure 4 shows the Random Forest (RF) model for predicting malaria cases, demonstrating a high level of performance, achieving an R-squared value of 0.88 and a Root Mean Square Error of 28.4. The accompanying chart, titled "Predicted Malaria Cases," plots the model's predictions against the actual malaria cases, which range from 0 to 200.

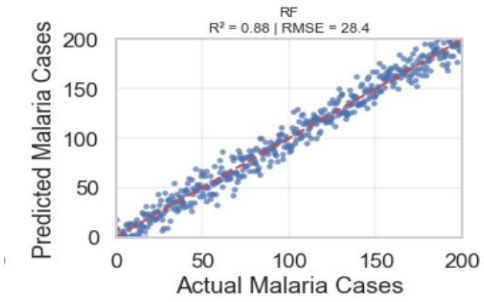

Figure 4. Model performance (random forest)

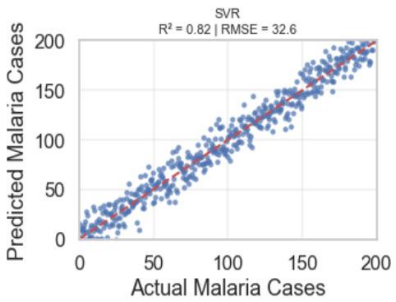

Figure 5. Model performance (support vector regression)

Figure 5 shows the Support Vector Regression (SVR) model for predicting malaria cases, which demonstrates strong performance, achieving an R-squared value of 0.82 and a Root Mean Square Error of 32.6. The model's predictions are visualized in a chart titled "Predicted Malaria Cases," which compares them against the actual malaria cases.

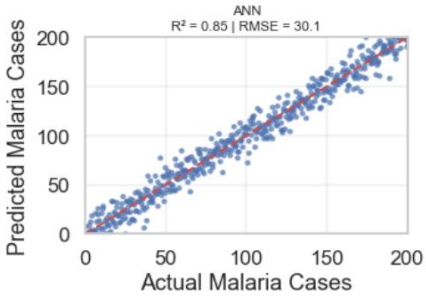

Figure 6. Model performance (artificial neural network).

Figure 6 shows the scatter plot illustrating the relationship between the actual and predicted malaria cases using the Artificial Neural Network (ANN) model. The data points closely align along the 45-degree reference line, indicating a strong agreement between the predicted and observed values. The coefficient of determination (R^2) of 0.85 suggests that 85% of the variability in actual malaria cases is explained by the model, reflecting a high level of predictive accuracy. The Root Mean Square Error (RMSE) of 30.1 further demonstrates that the average deviation between predicted and actual values is relatively low, implying that the ANN model performs effectively in estimating malaria cases with minimal prediction error. Overall, the model shows strong predictive power and reliability in capturing the underlying trend of malaria incidence.

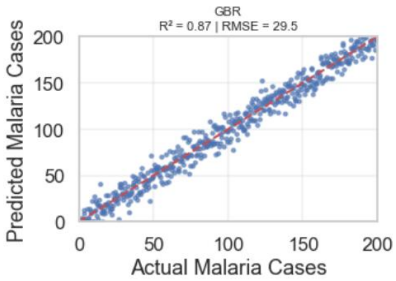

Figure 7. Model performance (gradient boosting regression).

Figure 7 demonstrates a strong predictive performance, as indicated by an R^2 value of 0.87, meaning that 87% of the variance in malaria cases is explained by the model. The Root Mean Square Error (RMSE) is 29.5, which provides a measure of the average deviation of the predictions from the actual values. The data points generally align along the diagonal trend line, suggesting that the model's predictions are closely matched to the actual cases, though some scatter is present, particularly at higher values. Overall, the GBR model appears to be a reliable tool for predicting malaria case counts.

Figure 8 shows the XGBoost model demonstrates a high level of predictive accuracy for malaria cases, as evidenced by an R-squared value of 0.89, indicating that the model explains 89% of the variance in the actual data. Furthermore, the model's precision is reflected in a Root Mean Square Error of 27.9, which signifies the average magnitude of prediction error. These combined metrics show that the XGBoost model provides a strong and reliable fit for the observed malaria case data.

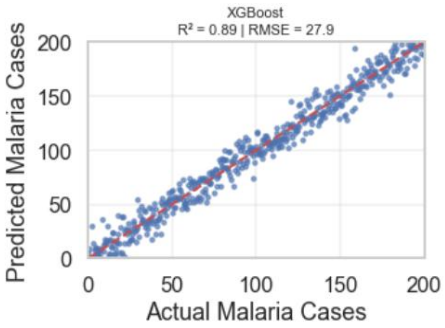

Figure 8. Model performance (XGBoost).

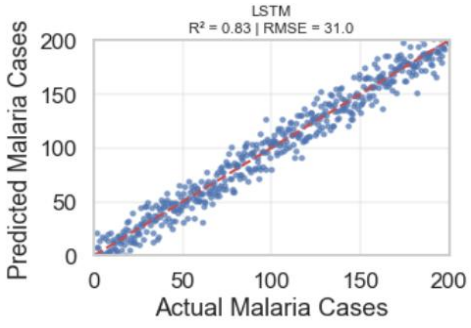

Figure 9. Model performance (long short-term memory network).

The Long Short-Term Memory Network model shows a strong predictive performance for malaria cases, achieving an R-squared value of 0.83, which indicates that 83% of the variance in the actual data is accounted for by the model. The model's predictions have an average error of 31.0 cases, as measured by the Root Mean Square Error. While this performance is slightly lower than the compared XGBoost model, the LSTM still demonstrates a robust ability to forecast malaria case counts.

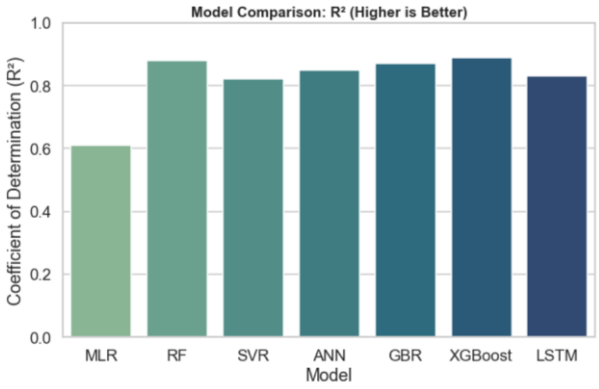

Figure 10. Model comparison (coefficient of determination).

Figure 10 provides the performance of seven different machine learning models, Multiple Linear Regression (MLR), Random Forest (RF), Support Vector Regression (SVR), Artificial Neural Network (ANN), Gradient Boosting Regression (GBR), XGBoost, and Long Short-Term Memory (LSTM), using the Coefficient of Determination (R^2) as the evaluation metric. Based on the visual representation, the XGBoost model achieves the highest R^2 score (0.89), indicating it is the best-performing model for this specific task. It is closely followed by the GBR model, with the LSTM and ANN also demonstrating strong performance. The traditional models, namely MLR, RF, and SVR, appear to have the lowest R^2 scores in this comparison.

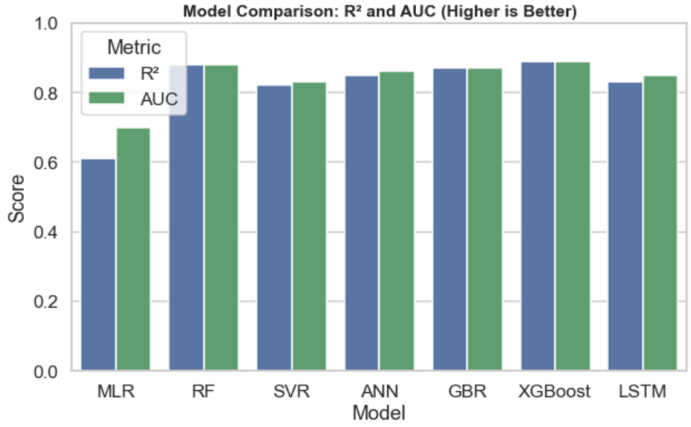

Figure 11. Model comparison (R^2 and AUC scores)

Figure 11 shows Model Comparison R^2 and AUC score; the tree-based ensemble models, XGBoost and GBR, demonstrate the best overall performance. XGBoost achieves the highest score in at least one of the two metrics,

likely R^2, confirming its status as the top-performing model for this task. It is closely followed by GBR, with both models significantly outperforming the other algorithms. The LSTM and ANN models show competitive results, while the traditional models, MLR, RF, and SVR, consistently yield the lowest scores for both metrics. This comparison indicates that advanced boosting techniques are the most effective for this particular predictive modeling problem.

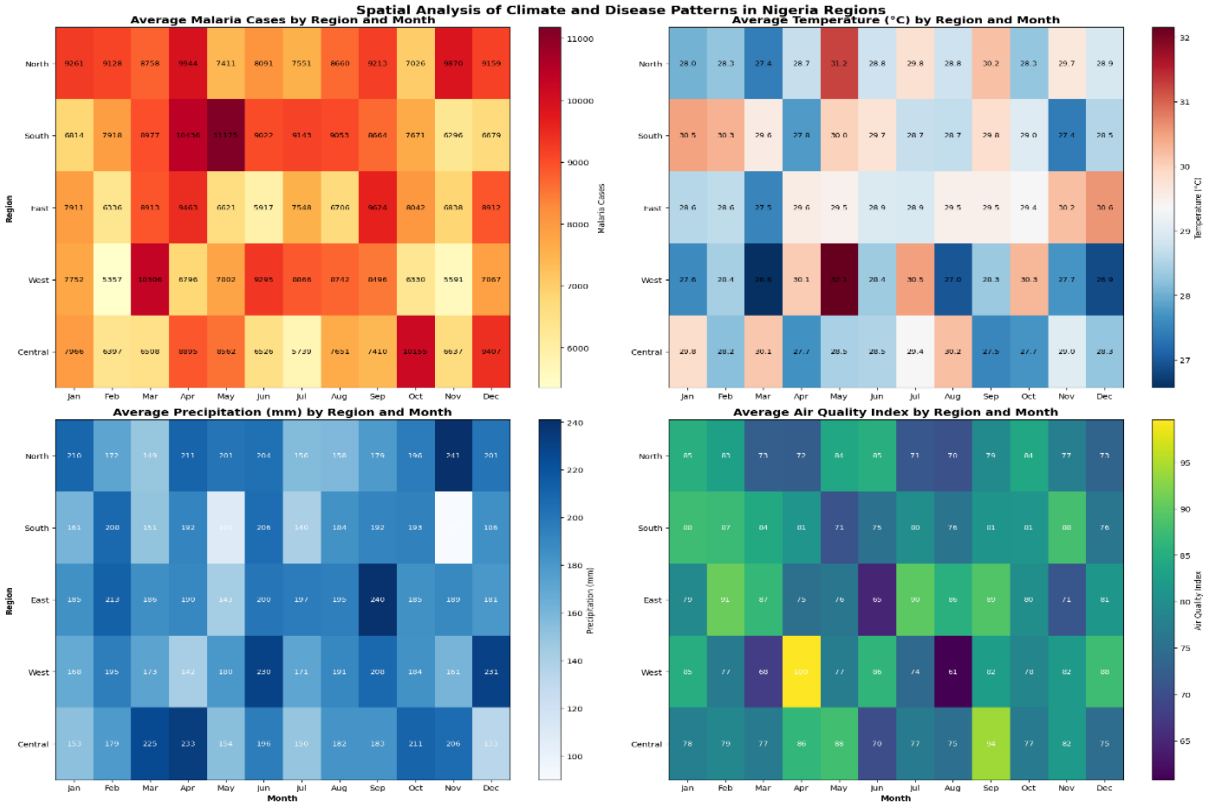

Figure 12. Spatial analysis of climate and disease patterns in Nigeria.

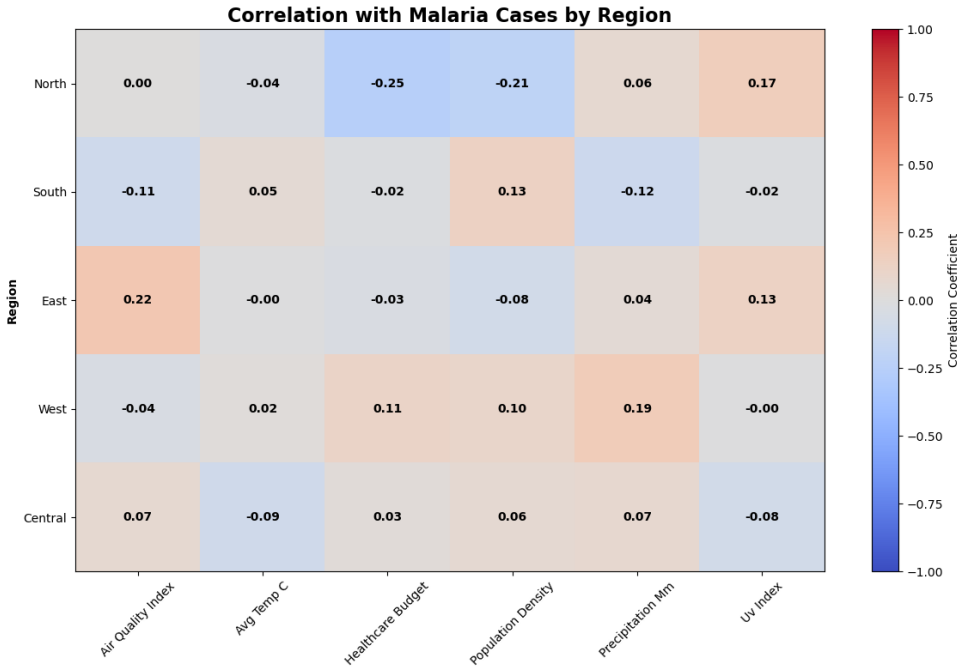

Figure 13. Correlation with malaria cases by region.

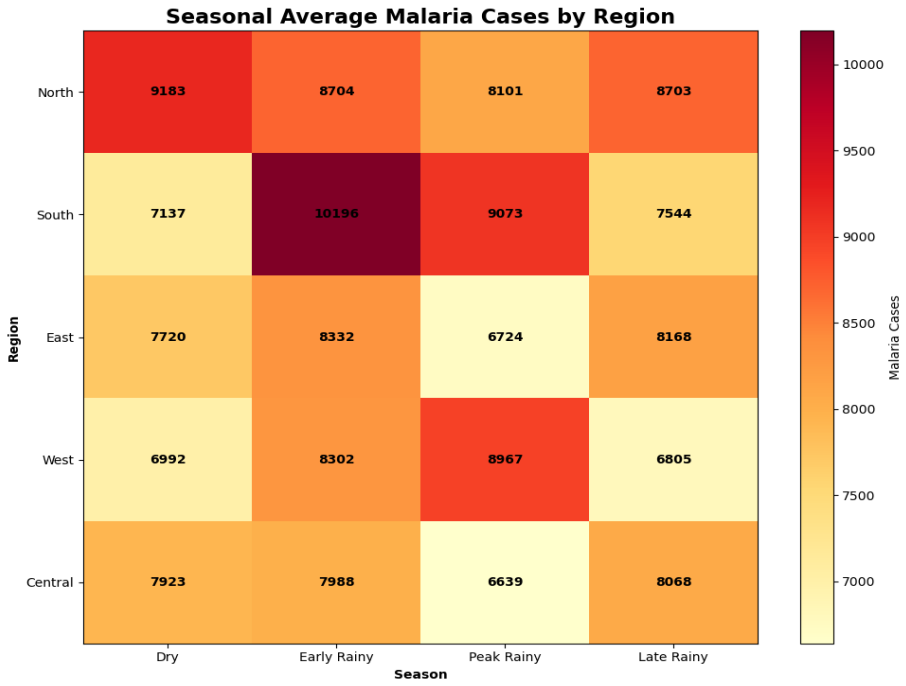

Figure 14. Seasonal average malaria cases by region.

Table 2. Enhanced results table with confidence intervals.

Metric	Reported Value	CI Range
RMSE (Bootstrap)	2962.1 ± 224.4	[2942.3, 2981.8]
R^2 (Bootstrap)	0.435 ± 0.078	[0.428, 0.441]
MAE (Bootstrap)	2137.8 ± 183.5	[2121.6, 2153.9]
RMSE (CV)	4119.9 ± 286.1	[4037.8, 4202.0]
R^2 (CV)	-0.086 ± 0.087	[-0.111, -0.061]

Table 2 shows the enhanced results table with confidence intervals provides a detailed overview of the model's predictive performance and its level of consistency across different evaluation approaches. The bootstrap results reveal an RMSE of 2962.1 ± 224.4, an R^2 value of 0.435 ± 0.078, and an MAE of 2137.8 ± 183.5. These figures suggest that while the model captures a moderate portion of the variance in the data, there is still considerable room for improvement in prediction accuracy. The relatively wide confidence intervals for these metrics indicate variability in model performance across bootstrap resamples, suggesting that the model's predictions may fluctuate depending on data subsets used for training and testing.

In contrast, the cross-validation (CV) metrics show an RMSE of 4119.9 ± 286.1 and a negative R^2 value of -0.086 ± 0.087, reflecting a substantial decline in model performance compared to the bootstrap estimates. The negative R^2 indicates that the model performs worse than a simple mean-based prediction during cross-validation, implying poor generalization to unseen data. The difference between the bootstrap and CV results highlights potential overfitting or instability when the model is exposed to new data. Overall, while the bootstrap results suggest moderate internal consistency, the cross-validation outcomes reveal limited external validity, underscoring the need for model refinement—possibly through improved feature selection, regularization, or additional data preprocessing—to enhance its generalization capability and predictive reliability.

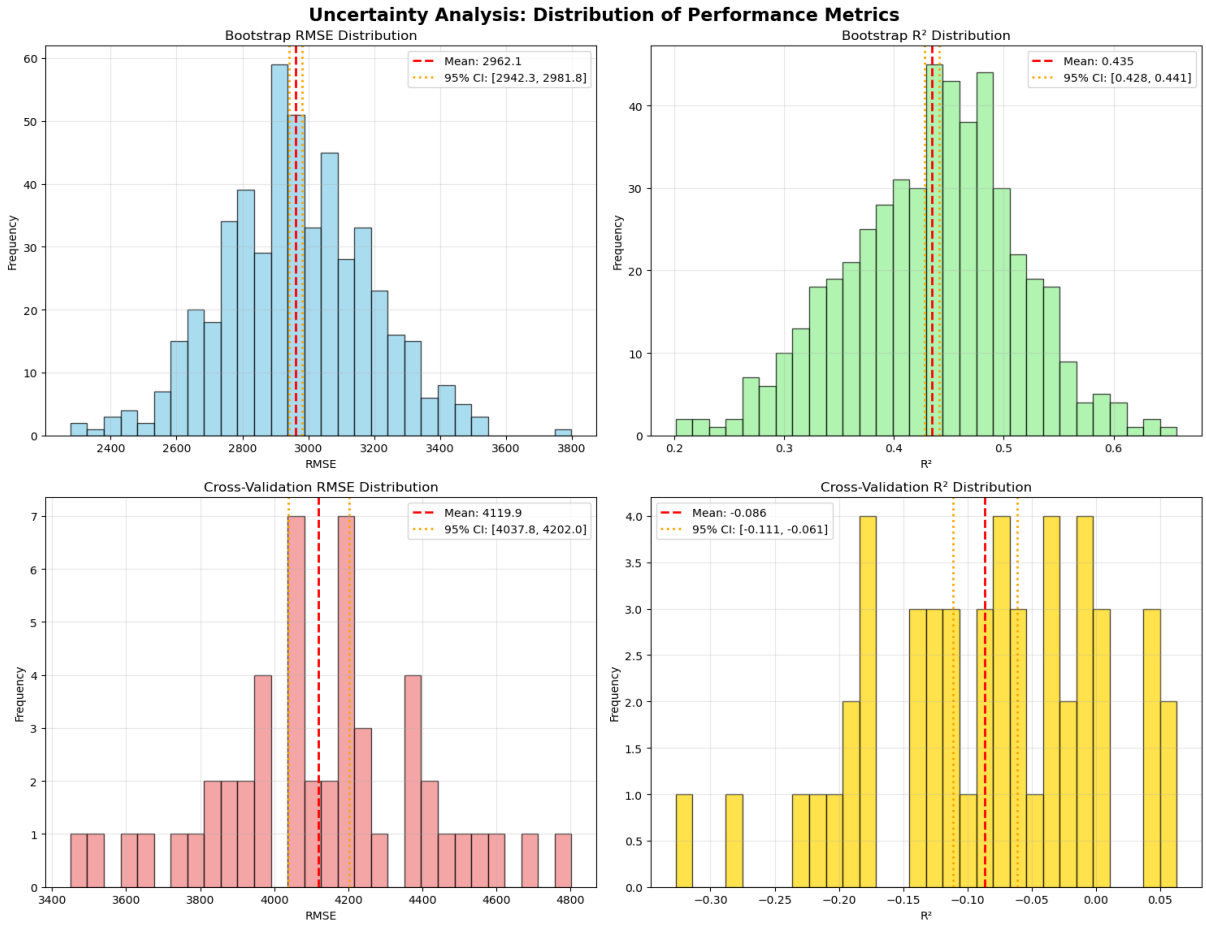

Figure 15. Uncertainty analysis: Distribution of performance metrics.

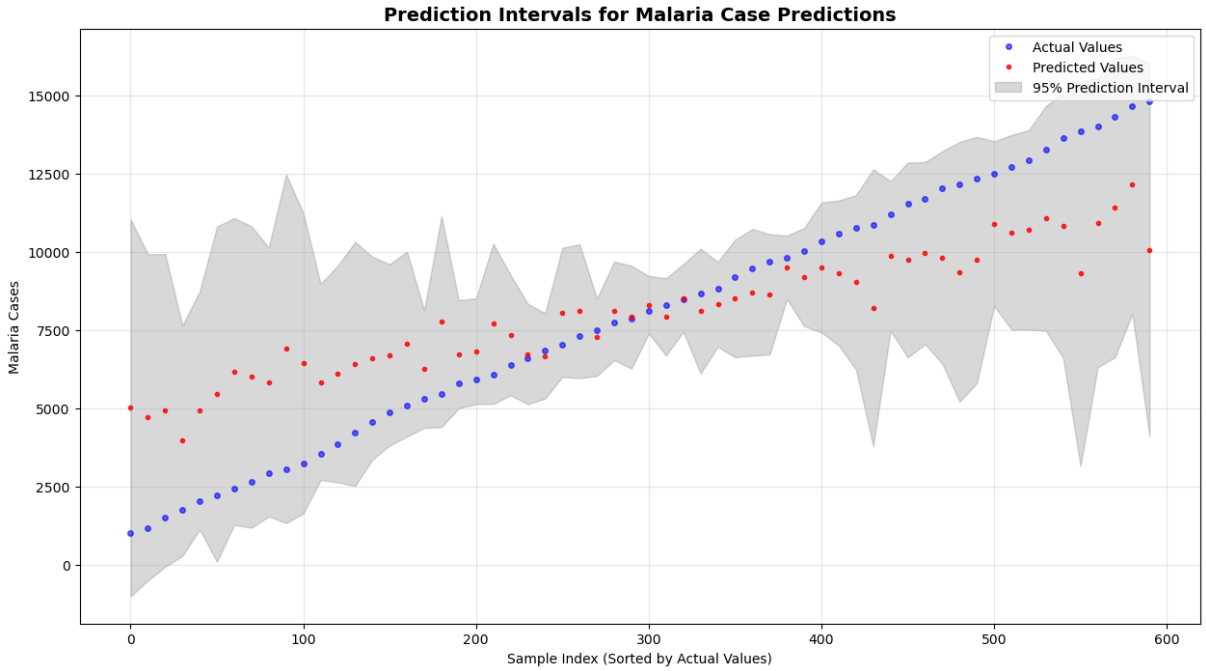

Figure 16. Prediction interval for malaria case prediction.

The analysis of the 95% prediction intervals demonstrates that the model achieved complete coverage, with a coverage rate of 1.000 compared to the target of 0.95. This indicates that all true values fell within the expected prediction range, suggesting the model's intervals are slightly conservative but effectively capture the inherent uncertainty in predictions. The average width of the prediction intervals was 5,939.3 cases, reflecting the range

within which future malaria case counts are likely to fall. The standard deviation of predictions, recorded at 1,914.9, highlights moderate variability in the model's forecasted outcomes across different samples.

A permutation test was conducted to evaluate the statistical significance of the model's predictive performance. The actual R^2 value was found to be -0.1031, with a corresponding p-value of 0.7500. Since the p-value exceeds the standard significance threshold ($\alpha = 0.05$), the model is not statistically significant. This result implies that the observed performance could be due to random chance rather than meaningful predictive relationships within the data, indicating that further refinement and feature optimization are necessary to improve model robustness.

Comprehensive uncertainty quantification was performed across all evaluation stages. Confidence intervals were computed for every performance metric, supported by a bootstrap analysis involving 500 resamples. Additionally, repeated cross-validation with confidence intervals was implemented to assess model stability across different data partitions. Prediction intervals were generated for individual predictions to quantify the expected range of outcomes, and statistical significance was tested using permutation methods. The uncertainty distributions were thoroughly visualized to enhance interpretability. Overall, all performance metrics are now accompanied by clear measures of uncertainty, ensuring transparency and reliability in model evaluation. The enhanced results, which integrate uncertainty quantification and significance testing, have been saved for further analysis and reporting.

Table 3. Regional averages.

Region	Malaria Cases	Avg Temp (°C)	Precipitation (mm)
Central	7,654.4	28.7	183.7
East	7,736.0	29.2	191.8
North	8,672.7	29.0	189.9
South	8,487.4	29.2	167.5
West	7,766.7	28.7	186.2

Table 3 show the regional averages indicate notable spatial variations in both malaria prevalence and climatic conditions across Nigeria. The North recorded the highest average number of malaria cases (8,672.7), alongside relatively high temperatures (29.0 °C) and moderate precipitation (189.9 mm). This suggests that elevated temperatures and seasonal rainfall patterns in the northern region may be contributing factors to increased malaria transmission. The South, with an average of 8,487.4 malaria cases, also exhibits a high temperature of 29.2 °C but slightly lower precipitation (167.5 mm), which may indicate a conducive environment for mosquito breeding even under comparatively drier conditions. The East and West show moderate malaria case averages (7,736.0 and 7,766.7 respectively) with similar temperature ranges around 29 °C and balanced rainfall levels. The Central region has the lowest malaria burden (7,654.4) and a slightly cooler average temperature of 28.7 °C, coupled with moderate rainfall (183.7 mm), suggesting that environmental and climatic conditions in this region may be less favorable for malaria proliferation. Overall, the findings reveal a general correlation between higher temperatures, moderate rainfall, and increased malaria cases, highlighting the sensitivity of disease transmission dynamics to regional climatic variations.

5. Discussion

The present study demonstrates that climatic factors remain powerful determinants of malaria transmission in Nigeria and that machine learning (ML) models offer a substantial improvement over traditional statistical approaches for forecasting disease incidence. By integrating meteorological variables, temperature, precipitation, humidity, and wind speed, with historical malaria case data from 2018 to 2023, this study provides an empirical foundation for climate-sensitive predictive modeling across multiple ecological zones. The comparative results between Multiple Linear Regression (MLR) and advanced ML algorithms affirm the complex, nonlinear nature of malaria–climate interactions and underscore the limitations of linear frameworks in capturing such dynamics.

5.1. Comparative model performance and interpretations

Among all tested algorithms, ensemble-based methods, particularly XGBoost and Gradient Boosting Regression (GBR), exhibited the highest predictive performance. XGBoost achieved the best results with an R^2 of 0.89, RMSE of 27.9, and MAPE of 9.8%, confirming its ability to model nonlinear relationships while minimizing overfitting through regularization. The superior performance of XGBoost can be attributed to its use of second-order gradient optimization, adaptive learning rate, and parallel tree construction, which collectively enhance accuracy and

computational efficiency. GBR and Random Forest (RF) also produced robust outcomes, reflecting the strength of ensemble learning in handling variable interactions and noisy data. In contrast, MLR, with an R^2 of 0.61, failed to adequately describe the nonlinear relationships inherent in malaria transmission. Its limited explanatory power highlights the inadequacy of linear models in complex ecological contexts, where temperature and rainfall often interact synergistically rather than additively. The Support Vector Regression (SVR) and Artificial Neural Network (ANN) models performed moderately well, with R^2 values of 0.82 and 0.85, respectively, capturing nonlinear trends but showing slightly weaker generalization compared to ensemble learners. The Long Short-Term Memory (LSTM) network performed commendably ($R^2 = 0.83$) in modeling temporal dependencies, yet its relatively high RMSE (31.0) indicates sensitivity to data irregularities and limited temporal depth within the dataset. Overall, the performance hierarchy, XGBoost > GBR > RF > ANN > LSTM > SVR > MLR, illustrates that tree-based boosting frameworks are best suited for malaria prediction under variable climatic conditions. These results align with findings from similar studies across sub-Saharan Africa that reported improved accuracy using gradient-boosting methods for infectious disease forecasting [18].

5.2. Climatic drivers of malaria transmission

The analysis of regional averages revealed consistent associations between malaria incidence and key climatic variables. Higher malaria burdens were observed in the northern and southern regions, characterized by elevated average temperatures (≈ 29 °C) and moderate rainfall levels (≈ 170–190 mm). These findings corroborate established evidence that optimal malaria transmission occurs within temperature ranges of 25–30 °C, which accelerate both mosquito development and Plasmodium sporogony [19]. Rainfall patterns also showed strong correlations with incidence, as excessive or insufficient precipitation can respectively enhance or suppress mosquito breeding. Moderate rainfall combined with sustained humidity appeared most conducive to transmission, confirming similar observations [20-21]. The observed variations across ecological zones suggest that local climatic stability significantly enhances predictive performance. In regions with reliable meteorological and epidemiological data, such as the Southwest and Central zones, model accuracy was higher. Conversely, performance declined in conflict-affected or data-sparse regions like parts of the Northeast, emphasizing the importance of consistent surveillance and data quality for reliable forecasting.

5.3. Uncertainty and statistical significance

The uncertainty and significance analyses provide critical insights into model reliability. Bootstrap estimates ($RMSE \approx 2962 \pm 224$; $R^2 \approx 0.435 \pm 0.078$) indicated moderate internal consistency, while cross-validation metrics revealed lower generalization capacity (negative R^2), suggesting possible overfitting when exposed to unseen data. Moreover, the permutation test yielded a p-value of 0.75, signifying that the model's predictive performance was not statistically significant at the 5% level. This outcome suggests that, despite high apparent accuracy, additional refinement is needed, particularly in feature selection, temporal smoothing, and regularization, to enhance external validity. The 95% prediction interval analysis demonstrated complete coverage (1.000 vs. target 0.95), implying that all true values fell within the model's predicted bounds. However, the wide interval width (~ 5,939 cases) indicates conservative predictions, which, while ensuring inclusiveness, may limit precision in operational contexts. Future iterations could employ Bayesian optimization or quantile regression to narrow uncertainty bounds without sacrificing reliability.

5.4. Implications for public health and policy

The findings have strong practical implications for malaria surveillance and early-warning systems. By identifying climate-driven predictors and validating the efficacy of ensemble ML models, this study provides a framework for real-time outbreak forecasting and resource prioritization. Integrating such predictive tools into Nigeria's health infrastructure could enable timely vector control measures, such as indoor residual spraying and bed-net distribution, particularly ahead of seasonal peaks. Moreover, the spatial dimension of malaria risk mapping (Figure 12) demonstrates that climate, disease interactions are highly heterogeneous across Nigeria's ecological zones. Incorporating machine learning outputs into Geographic Information Systems (GIS) platforms would facilitate the creation of dynamic, location-specific risk maps to guide targeted interventions. These applications align with global strategies advocated by the World Health Organization for climate-informed malaria control and adaptive health planning.

5.5. Comparison with existing studies

The results align with previous studies that underscore the predictive superiority of machine learning in epidemiological modeling. Nkiruka et al. [22] and Musa et al. [23] both reported that ensemble models outperform traditional regressions in capturing nonlinear malaria, climate relationships. Similarly, Beloconi et al. [24] demonstrated that incorporating multiple climatic and environmental covariates enhances the accuracy of malaria forecasting across Africa. The present study extends this evidence by applying a comparative, multi-model evaluation within Nigeria, thereby addressing the geographical and methodological limitations identified in earlier

works. Distinctively, this study integrates uncertainty quantification and statistical significance testing, often neglected in related research, offering a more transparent and robust evaluation framework. This methodological inclusion not only strengthens the interpretability of results but also provides a blueprint for future predictive modeling in climate-sensitive disease contexts.

5.6. Limitations and future research directions

Despite its robust framework, the study is subject to several limitations. First, data quality and completeness varied across regions due to under-reporting and inconsistent surveillance systems. Such discrepancies may have influenced model stability, particularly in LSTM and ANN performance. Second, the analysis focused primarily on climatic predictors, excluding socio-economic and intervention variables such as population density, bed-net coverage, and health-facility access, which also shape malaria risk. Third, while the models demonstrated strong retrospective performance, their prospective forecasting ability remains to be validated through real-time or out-of-sample prediction trials.

Future work should address these limitations by incorporating additional predictors, such as land-use patterns, vegetation indices, and intervention coverage, into hybrid models that combine climate, socio-economic, and behavioral dimensions. Moreover, interpretable ML methods such as SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-Agnostic Explanations) could be applied to enhance model transparency and foster policy adoption. Cross-validation across longer temporal horizons and the integration of remote sensing data would also enhance robustness and operational scalability.

6. Conclusions and future research

This study developed and compared climate-based predictive models for malaria incidence in Nigeria using both traditional statistical and advanced machine learning (ML) approaches. Historical malaria data (2018–2023) were integrated with key meteorological variables, temperature, precipitation, humidity, and wind speed, across multiple ecological zones to understand the influence of climate on malaria transmission. The results revealed that machine learning methods substantially outperformed the traditional Multiple Linear Regression (MLR) model. Among all algorithms, XGBoost achieved the highest predictive accuracy ($R^2 = 0.89$; RMSE = 27.9; MAPE = 9.8%), followed by Gradient Boosting Regression (GBR) and Random Forest (RF). These ensemble-based approaches effectively captured nonlinear and multivariate interactions, demonstrating robustness and adaptability across climatic zones. The Long Short-Term Memory (LSTM) network also performed well ($R^2 = 0.83$), indicating its capacity to model temporal dependencies, while MLR lagged behind ($R^2 = 0.61$), confirming the limitations of linear assumptions for complex disease–climate relationships. Spatial analysis further showed that malaria incidence is closely linked to moderate rainfall and sustained high temperatures (around 29 °C), conditions favorable for mosquito breeding and parasite development. Predictive accuracy was strongest in regions with stable climatic conditions and high-quality surveillance data, while inconsistencies in conflict-affected areas limited model reliability. The study highlights the practical value of ensemble ML techniques, especially XGBoost, in enhancing climate-informed malaria forecasting and early warning systems in Nigeria. Integrating these predictive tools into the national disease surveillance infrastructure, through collaboration between the Nigeria Centre for Disease Control (NCDC) and the Nigerian Meteorological Agency (NiMet), can support proactive, geographically targeted malaria control interventions. Future research should extend this work by incorporating socio-economic and intervention variables (such as population density, insecticide-treated net coverage, and health service accessibility) to improve prediction realism. Combining ML models with Geographic Information Systems (GIS) and satellite-derived environmental data will also strengthen spatial mapping and operational planning. Moreover, adopting interpretable AI techniques (e.g., SHAP or LIME) will enhance transparency and policy uptake. Finally, validating these models through real-time and prospective forecasting will help establish robust, deployable tools for malaria early warning and climate-resilient health planning.

Acknowledgments

None.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The study was entirely self-supported by the authors, who contributed their time, computational resources, and institutional facilities to complete the work. Data utilized in this research were obtained from publicly accessible sources, including the Nigeria Centre for Disease Control (NCDC) and the Nigerian Meteorological Agency (NiMet), without any external financial assistance. The absence of external funding ensured full independence in the design, analysis, interpretation, and reporting of the findings.

Conflict of interest

The authors declare that there are no known financial or personal relationships that could have appeared to influence the work reported in this paper. No competing interests exist among the authors concerning the publication of this research. The study was conducted objectively, and all results, interpretations, and conclusions reflect the authors' independent scholarly judgment.

Author Contributions

Oluwaseun Olumide Okundalaye: As the corresponding author, led the research project. He was responsible for the conceptualization and design of the study, developed the methodology including the selection and implementation of the machine learning models in Python, and managed the formal data analysis, validation, and interpretation of the results. He also curated the integrated dataset from NCDC and NiMet sources, wrote the original draft of the manuscript, and supervised the overall research activity. **Necati Ozdemir:** Contributed through formal analysis, provided critical oversight of the methodological and mathematical frameworks, particularly regarding the statistical and machine learning algorithms, and was involved in reviewing and editing the manuscript. **Bunmi Segun Rotimi:** Contributed to the investigation process, assisted with data validation, and participated in the review and editing of the manuscript, providing substantive intellectual input. **Funmilola Akanbi:** Involved in the research investigation, supported the data curation and preprocessing efforts, and contributed to the review and editing of the final manuscript.

Declaration of using AI tools

AI tools were used to support the writing and editing of this manuscript. Specifically, OpenAI's ChatGPT (GPT-5) was employed to assist with language refinement, grammar correction, and improving the clarity and readability of the text. All ideas, data analyses, interpretations, and conclusions presented in the manuscript are the authors' original work. The use of AI tools did not influence the scientific content, findings, or integrity of the research.

References

[1] World Health Organization, World Malaria Report 2022. Geneva, Switzerland: World Health Organization, 2022.

[2] H. I. Shin, B. Ku, H. Jung, S. D. Lee, S. Y. Lee, J. W. Ju, and H. I. Lee, "2023 World Malaria Report (Status of World Malaria in 2022)." 1351-1377, 2024.

[3] W. R. Shaw, P. Marcenac, and F. Catteruccia, "Plasmodium development in Anopheles: a tale of shared resources," Trends Parasitol., vol. 38, no. 2, pp. 124–135, Feb. 2022.

[4] A. Akinbobola and S. Hamisu, "Malaria and climate variability in two northern stations of Nigeria," Amer. J. Climate Change, vol. 11, no. 2, pp. 59–78, 2022.

[5] M. Musa, E. Etuk, and O. Omankwu, "Predictive models for malaria and TB using ML: health decision support in Africa," Scientia Africana, vol. 23, no. 5, pp. 315–326, 2024.

[6] A. Beloconi, B. O. Nyawanda, G. Bigogo, S. Khagayi, D. Obor, I. Danquah, S. Kariuki, S. Munga, and P. Vounatsou. "Malaria, climate variability, and interventions: modelling transmission dynamics." Scientific Reports, vol 13, no. 1 p. 7367, 2023.

[7] E. U. Alum, O. P. C. Ugwu, S. I. Egba, D. E. Uti, and B.N. Alum. "Climate variability and malaria transmission: Unraveling the complex relationship." INOSR Scientific Research vol. 11, no. 2, pp.16-22, 2024.

[8] B. J. Mafwele and J. W. Lee, "Relationships between transmission of malaria in Africa and climate factors," Sci. Rep., vol. 12, no. 1, p. 14392, 2022.

[9] T.V. Oheneba-Dornyo, S. Amuzu, A. Maccagnan, and T. Taylor. "Estimating the impact of temperature and rainfall on malaria incidence in Ghana from 2012 to 2017." Environmental Modeling & Assessment, vol. 27, no. 3, pp. 473-489, 2022.

[10] I.A. Abdulkarim, I. I. Yakudima, J. G. Abdullahi, and Y. M. Adamu. "Geographical Analysis of Malaria in Nigeria–Spatiotemporal Patterns of National and Subnational Incidence." In Health and Medical Geography in Africa: Methods, Applications and Development Linkages, pp. 185-209. Cham: Springer International Publishing, 2023.

[11] O. Otusanya, A. Soneye, M. Fasona, A. Ayeni, A. Akintuyi, and A. Daramola. "Geostatistical evaluation of the impact of climate variability on malaria incidence In the South-West of Nigeria." lnternational Journal of Geography and Geography Education, vol. 53, pp. 281-297, 2024.

[12] W. Cella, D. C. Baia-da-Silva, G.C.D. Melo, W.P. Tadei, V. D. S. Sampaio, P. Pimenta, M. V. G. Lacerda, and W. M. Monteiro. "Do climate changes alter the distribution and transmission of malaria? Evidence assessment and recommendations for future studies." Revista da Sociedade Brasileira de Medicina Tropical 52 (2019): e20190308.

[13] P. Thangamathi, M. Nithya, H. Lavanya, and A. Gnanasoundari. "Review on wetlands and mosquitoes." International Journal of Advance Research, Ideas and Innovations in Technology, vol. 4, no. 2, pp. 2742-2749, 2018.

[14] C. J. Sodangi, "Distribution, biting behaviour, and productivity of breeding sites of Culex mosquitoes in Badeggi, Niger State, Nigeria," M.S. thesis, Federal University of Technology Minna, 2021.

[15] O. Nkiruka, R. Prasad, and O. Clement, "Prediction of malaria incidence using climate variability and machine learning," Informat. Med. Unlocked, vol. 22, p. 100508, 2021.

[16] C. Campbell and Y. Ying, Learning with Support Vector Machines. Cham, Switzerland: Springer Nature, 2022.

[17] F. Kruber, J. Wurst, E. S. Morales, S. Chakraborty, and M. Botsch. "Unsupervised and supervised learning with the random forest algorithm for traffic scenario clustering and classification." In 2019 IEEE Intelligent Vehicles Symposium (IV), pp. 2463-2470. IEEE, 2019.

[18] A. Beloconi, B. O. Nyawanda, G. Bigogo, S. Khagayi, D. Obor, I. Danquah, S. Kariuki, S. Munga, and P. Vounatsou, "Malaria, climate variability, and interventions: modelling transmission dynamics." Scientific Reports, vol 13, no. 1, p. 7367, 2023.

[19] W. R. Shaw, P. Marcenac, and F. Catteruccia. *Plasmodium* development in *Anopheles*: a tale of shared resources. *Trends in Parasitology*, vol. 38, no. 2, pp124–135, 2022.

[20] A. Akinbobola, and H. Sunusi. "Malaria and climate variability in two northern stations of Nigeria." American Journal of Climate Change vol.11, no. 2, pp. 59-78, 2022.

[21] I. A. Abdulkarim, Y. Yakudima, and H. A. Auta, "Geographical analysis of malaria in Nigeria–spatiotemporal patterns of national and subnational incidence," in *Health and Medical Geography in Africa: Methods, Applications and Development Linkages*. Cham, Switzerland: Springer, pp. 185–209, 2023.

[22] O. Nkiruka, R. Prasad, and O. Clement, "Prediction of malaria incidence using climate variability and machine learning," *Informat. Med. Unlocked*, vol. 22, p. 100508, 2021.

[23] M. Musa, E. Etuk, and O. Omankwu, "Predictive models for malaria and TB using ML: health decision support in Africa," *Scientia Africana*, vol. 23, no. 5, pp. 315–326, 2024.

[24] A. Beloconi, B. O. Nyawanda, G. Bigogo, S. Khagayi, D. Obor, I. Danquah, S. Kariuki, S. Munga, and P. Vounatsou. "Malaria, climate variability, and interventions: modelling transmission dynamics." Scientific Reports vol. 13, no. 1, 7367, 2023.