

***Machine Learning-Based Prediction of Malaria Risk Among Rural Households in
Kebbi State, Nigeria*****Abubakar Danjuma Bundaram¹ and Sabiu Lawali Tsafe^{2*}**¹Department of Public Health, Federal University of Medical and Health Sciences Funtua, Katsina, Nigeria²Department of Computer Science, Federal University Birnin Kebbi, Nigeria

Citation: Abubakar Danjuma Bundaram, Sabiu Lawali Tsafe (2026) Machine Learning-Based Prediction of Malaria Risk Among Rural Households in Kebbi State, Nigeria. *Int J. of Gastro Med & Res* 1(3), 1-17.
WMJ-IJGMR-26-110

Abstract

Malaria remains a major public health burden in rural sub-Saharan Africa, where environmental and socioeconomic factors interact in complex, non-linear ways that traditional statistical models often fail to capture. This study examines machine learning-based prediction of malaria risk among 200 rural households in Kebbi State, Nigeria, integrating environmental exposure data (including proximity to stagnant water) with socioeconomic and behavioural survey indicators collected through a door-to-door household survey. Ensemble learning techniques, namely Random Forest and XGBoost, were applied alongside Decision Tree, Support Vector Machine, and logistic regression to identify the most significant predictors of malaria infection and to compare predictive performance across algorithms. Model performance was evaluated using accuracy, precision, recall, F1-score, and AUC-ROC on a held-out 30% test set following stratified 10-fold cross-validation hyperparameter tuning. Logistic regression and XGBoost achieved the strongest discrimination (AUC = 0.90 and 0.90, respectively), with Random Forest close behind (AUC = 0.89); consistent insecticide-treated net (ITN) use emerged as the single strongest predictor across all methods, followed by prior malaria history, stagnant water near the house, and overgrown bushes around the house. These findings indicate that machine learning approaches, including but not limited to ensemble methods, can support household-level malaria risk identification in resource-limited rural settings, offering actionable insight for targeted vector-control and behavioural interventions.

***Corresponding author:** Sabiu Lawali Tsafe, Department of Computer Science, Federal University Birnin Kebbi, Nigeria.

Submitted: 02.08.2026**Accepted:** 11.08.2026**Published:** 07.09.2026

Keywords: Malaria Risk Prediction, Machine Learning, Rural Households, Kebbi State, Nigeria, Random Forest, XGBoost, Logistic Regression

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area Under the Curve
CI	Confidence Interval
DHS	Demographic and Health Survey
DVWA	Damn Vulnerable Web Application
IRS	Indoor Residual Spraying
ITN	Insecticide-Treated Net
LGA	Local Government Area
MIS	Malaria Indicator Survey
OR	Odds Ratio
RDT	Rapid Diagnostic Test
ROC	Receiver Operating Characteristic
SD	Standard Deviation
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine

Introduction

Malaria remains a persistent public health challenge in sub-Saharan Africa, disproportionately affecting rural populations where environmental and socioeconomic vulnerabilities converge [1]. In regions such as Kebbi State, proximity to stagnant water bodies has been linked to exacerbated transmission dynamics, necessitating analytical frameworks capable of capturing complex, non-linear risk patterns [2]. By integrating granular socioeconomic and environmental survey data with machine learning algorithms, this study seeks to improve the accuracy of local malaria risk prediction models beyond the capabilities of traditional statistical methods [3,4]. While traditional logistic regression provides insight into individual risk factors, ensemble and other machine-learning approaches can in some settings resolve more nuanced, multi-faceted interactions between household environmental exposure and behavioural prophylaxis than a single linear model [5]. Consequently, this study employs several supervised classifiers — including Random Forest and XGBoost — to identify predictors of malaria infection in Kebbi State and to benchmark them against an interpretable logistic-regression baseline [6]. The study leverages a structured dataset of 200 rural households, collected through a door-to-door survey administered by trained enumerators, to train and validate these models, using accuracy, precision, recall, F1-score, and AUC-ROC to assess predictive

performance and clinical utility.

Literature Review

Recent literature underscores a growing interest in applying machine learning to malaria surveillance, particularly in West Africa where predictive analytics are increasingly explored for vector surveillance and outbreak forecasting [7]. Researchers have increasingly utilised gradient boosting and ensemble classifiers to identify high-risk drivers, such as household environmental conditions and socioeconomic status, which simpler logistic models may not capture as fully [8]. By aggregating multiple decision-based learners, ensemble approaches can help resolve non-linear dependencies between localised environmental hazards and malaria incidence [9,10]. Studies in resource-limited settings further suggest that integrating demographic features with clinical and behavioural variables can enhance model robustness, supporting more precise identification of at-risk populations [11,12].

Several studies report machine learning algorithms providing strong discrimination relative to conventional regression techniques when classifying household-level malaria risk profiles [13,14], and algorithms such as Random Forest and XGBoost have shown efficacy in handling the high-dimensional, heterogeneous data characteristic of rural epidemiological surveys [15]. Other work has integrated susceptibility, vulnerability, and adaptive-capacity factors into model design, which can be useful for informing health policy in endemic regions [16]. Comparable ensemble approaches based on XGBoost, Random Forest, and ecological-niche modelling have also been deployed for *Plasmodium knowlesi* transmission-risk mapping in Malaysia, where combining the outputs of multiple models produced stronger discrimination than any single model in isolation [14].

The Burden of Malaria in Nigeria and the West African Context

Nigeria continues to dominate the global malaria landscape, contributing an estimated 27% of all malaria cases and 31% of all malaria deaths worldwide, making it the most heavily burdened country on the continent [17,18]. Sub-Saharan Africa as a region accounts for approximately 90% of global malaria mortality, reflecting the convergence of intense perennial transmission, weak health-system infrastructure, and entrenched socioeconomic vulnerability [19]. Within

Nigeria, the 2018 Demographic and Health Survey found that 35.5% of children aged 6–59 months tested positive for malaria by rapid diagnostic test, with sub-national prevalence showing extreme spatial heterogeneity: Kebbi State recorded the highest burden at 77.6% (95% CI: 70.2–83.5), followed by Katsina State at 55.5%, while the Federal Capital Territory recorded only 29.6% [17]. Children in rural residential locations were found to be 2.12 times more likely (adjusted odds ratio = 2.12, 95% CI: 1.75–2.57) to test positive than their urban counterparts [17].

A community-based study of 551 Hausa participants across five Local Government Areas in neighbouring Kano State found an overall *Plasmodium falciparum* prevalence of 60.6%, with multivariate analysis identifying age ≥ 12 years, low household income, non-use of insecticide-treated nets, and absence of household toilet facilities as independent risk factors [20]. Within Kebbi State specifically, a hospital-based RDT survey of 300 consenting individuals in Birnin Kebbi reported a prevalence of 20.7%, with the highest infection rates among those aged 11–20 years (31.4%) and the lowest among children aged 0–10 years (13.3%), though neither age nor gender differences reached statistical significance [21]. A companion socioeconomic study in Birnin Kebbi further documented that illiterate participants recorded significantly higher malaria rates (28.0% vs. 18.0%, $p = 0.042$) than their literate counterparts, with farmers also bearing a heavier burden (29.5%) compared with civil servants (16.8%) and business operators (20.6%) [22]. Taken together, these findings situate the present study within an epidemiological landscape characterised by extreme spatial heterogeneity, entrenched rural vulnerability, and a constellation of behavioral, environmental, and socioeconomic risk drivers.

Determinants of Household-Level Malaria Risk

The household is an analytically useful unit of analysis for malaria risk prediction because transmission events occur within the peridomestic environment, shaped by household-level behavioural, socioeconomic, and spatial factors [23]. Across diverse African settings, several consistent household-level determinants have emerged: large household size and overcrowding can increase odds of infection by reducing intra-household ITN access and raising the probability of intra-household transmission [24]. In the Democratic Republic of the Congo, household-level analysis

similarly found that socioeconomic vulnerability and weak vector-control coverage remained principal modifiable drivers of malaria burden, even where ITN ownership was relatively high [25]. A Bauchi State, Nigeria, study of 300 patients found that rural residency (29.67%), low-income status (20.67%), and poor housing and environmental conditions all disproportionately elevated malaria risk, with recent fever history (within two weeks) emerging as the only statistically significant predictor ($p = 0.006$) in the adjusted regression [26]. These findings reinforce the view that household malaria status is multifactorial and shaped by intersecting structural vulnerabilities, supporting analyses capable of accommodating that multi-dimensionality [3,27].

Advances in Machine-Learning Modelling for Malaria Surveillance

Parallel to the evolution of epidemiological understanding, the methodological frontier in malaria surveillance has increasingly incorporated machine-learning inference. In Nigeria, Odu and colleagues found that an XGBoost-based classifier integrating 28 years of climate variability data across six Sub-Saharan African countries outperformed competing algorithms in classifying malaria incidence, with precipitation, temperature, and surface radiation identified as dominant non-seasonal predictors [19]. A separate Nigeria-focused climate-based modelling study comparing seven algorithms found that XGBoost produced the highest predictive accuracy for malaria incidence ($R^2 = 0.89$, RMSE = 27.9, MAPE = 9.8%), followed closely by Gradient Boosting Regression and Random Forest, while a traditional multiple linear regression baseline achieved only $R^2 = 0.61$ [28]. Deep-learning extensions have further refined the predictive frontier: a satellite-derived dataset covering 2000–2017 found that a Long Short-Term Memory network achieved a notably lower error than the best-performing Transformer configuration for malaria-incidence prediction in Nigeria, underscoring the value of temporal sequential modelling for capturing seasonality and lagged environmental effects [29].

Nationally, a study integrating DHS/MIS biomarker datasets with 34 years of CHIRPS rainfall and MODIS-NDVI vegetation indices in a Random Forest framework achieved cross-year validation accuracies of 62.2%, 36.8%, and 40.2% across three survey years, with SHAP-based interpretability highlighting ITN

coverage, a three-year climate lag of rainfall/NDVI, temperature, and population density as dominant predictors of sub-national risk [8]. At the community level, a Nigerian childhood-malaria study using 2015–2020 DHS data found that Random Forest achieved an AUC of 0.89, outperforming a conventional logistic regression baseline (AUC = 0.82), while Gradient Boosting achieved AUC = 0.87 and a Decision Tree lagged at AUC = 0.78 — a pattern broadly consistent with literature favouring ensemble learners for infectious-disease prediction [3,1]. A separate Ethiopian study using PMA survey data on bed-net utilisation reported Random Forest AUC values around 0.89–0.91 alongside XGBoost AUC around 0.88, depending on optimisation [30]. Comparable ensemble approaches have also been deployed for *Plasmodium knowlesi* transmission-risk mapping in Malaysia [14]. Although such applications span heterogeneous geographies, they collectively suggest that ensemble learning, careful feature engineering, and rigorous cross-validation are valuable ingredients for translatable malaria-risk models in data-scarce, resource-constrained settings [3,11,12,28].

Justification of the Present Study

The evidence reviewed above establishes several converging motivations for the present study. First, Kebbi State is a documented national hotspot for malaria transmission, with the highest under-five prevalence recorded in the 2018 Nigeria DHS and persistent household-level risk factors that are themselves spatially clustered [22,17,2]. Second, risk is multi-causal and plausibly non-linear, encompassing environmental (stagnant water, vegetation, rainfall), behavioural, and socioeconomic (income, education, occupation) dimensions that can be difficult for a single linear model to capture fully [23,26,31]. Third, ensemble machine-learning classifiers have, in several comparable African contexts, offered discrimination at least on par with — and sometimes better than — logistic regression [3,1,27,28]. Accordingly, the present study applies a multi-algorithm comparison to cross-sectional household-level data collected from rural Kebbi State communities, with the dual objective of (a) benchmarking machine-learning discrimination against an interpretable logistic-regression baseline and (b) extracting feature-importance rankings that can inform targeted, household-level intervention design within the State's malaria elimination framework [5,15,10].

Materials and Methods

Study Design

This study adopted a cross-sectional, community-based quantitative design employing primary household-level survey data to develop and evaluate machine learning classification models for predicting malaria risk among rural households in Kebbi State, Nigeria. A non-experimental, associational research approach was used, in which risk-factor exposure and household malaria status were measured concurrently without manipulation of independent variables [6]. This design is consistent with established practice in data-driven malaria epidemiology, where cross-sectional datasets support the fitting of supervised machine-learning algorithms capable of extrapolating risk to comparable populations [23].

Study Area

The study was conducted in Kebbi State, a north-western Nigerian state situated within the Sahel and Sudan savannah ecological zones. The state is characterised by a tropical climate with a single, intense rainy season (June–September) and a prolonged dry season, conditions that historically sustain seasonal *Anopheles* mosquito breeding and high malaria transmission intensity. Kebbi State also has a large rural population engaged in subsistence agriculture, and proximity to stagnant water bodies has been identified as a contributor to elevated mosquito density and malaria transmission risk in the region. The state also reports a poverty rate of approximately 72.0%, which can compromise household-level access to preventive measures such as insecticide-treated nets and timely diagnosis [2].

Five rural communities were purposively selected to represent distinct agro-ecological and socio-cultural micro-settings within the state:

S/N	Community	LGA Context	Households Surveyed
1	Kalgo	Kalgo LGA	40
2	Kardi	Birnin Kebbi LGA	40
3	Yamama	Kalgo LGA	40
4	Basaura	Jega LGA	40
5	Maidahini	Bunza LGA	40
	Total		200

Equal allocation across communities was employed to permit balanced sub-group comparisons in descriptive and bivariate analyses.

Study Population

The target population comprised rural households across the selected communities. The unit of analysis was the household, with the household head or a knowledgeable adult representative acting as the primary respondent for questionnaire administration.

Sample Size

The achieved sample size for this study was 200 households (40 per community). While no formal a priori power calculation was conducted, this sample size is comparable to similar machine-learning pilot studies conducted in resource-limited African settings, and is sufficient to support stratified cross-validation and the training of several classifiers without severe overfitting, particularly when the predictor space is moderate and a 70:30 hold-out or k-fold scheme is employed [3,12].

Sampling Technique

A multi-stage cluster sampling strategy was implemented:

Stage 1 - LGA Selection: Birnin kebbi, Bunza and Kalgo LGA were selected because of their documented malaria burden, their predominantly rural character, and their confluence of environmental risk factors (proximity to water bodies and seasonal flooding).

Stage2 - Community Selection: Five rural communities within the Birnin Kebb, Bunza and Kalgo were purposively identified to capture heterogeneity in environmental exposure and ethnic composition.

Stage 3 - Household Selection: Within each community, 40 households were selected using systematic random sampling from a community-generated household list. Where such a list was unavailable, a spin-the-pan technique was applied at a central location in each settlement. This procedure mirrors the multi-stage approach commonly used in Demographic and Health Survey-style malaria surveillance and supports representativeness at the community level while remaining operationally feasible in hard-to-reach rural settings [4,5].

Inclusion and Exclusion Criteria

Inclusion criteria:

Households that had been resident in the selected community for at least six months prior to data collection. Households with a consenting adult respondent (≥ 18 years) available at the time of visit.

Verbal or written informed consent provided. Exclusion criteria:

Households that had recently relocated into the community.

Households in which no eligible adult respondent was available after repeated visits.

Data Collection

Data were collected through a structured, interviewer-administered questionnaire delivered door-to-door by trained enumerators across the five selected communities. The questionnaire captured demographic, socioeconomic, environmental, behavioural, and health-related variables, summarised in the table below.

Domain	Variable	Type	Categories / Unit
Demographic	Age of respondent	Continuous	Years
	Gender	Categorical	Male / Female
	Education Level	Ordinal	None / Primary / Secondary / Tertiary
	Household Size	Discrete	Number of residents
Socioeconomic	Income Level	Ordinal	Low / Medium / High
Environmental	Stagnant water near house	Binary	Yes / No
	Bushes around house	Binary	Yes / No
	Water source	Categorical	Borehole / Well / River
	Drainage condition	Binary	Good / Poor
Behavioural	Consistent ITN usage	Binary	Yes / No
	Indoor Residual Spraying	Binary	Yes / No
	Outdoor night activities	Binary	Yes / No
Health	Previous malaria history (12 months)	Binary	Yes / No
	Distance to nearest health facility	Continuous	Kilometres
Outcome	Household malaria status	Binary	Positive / Negative

The outcome variable - household malaria status - was recorded as positive if a household member had been diagnosed with malaria (by RDT or microscopy) within the recall window used by the survey instrument, consistent with the recall-window approach frequently used in DHS/MIS-derived malaria datasets [27,8].

Data Preprocessing

Prior to model fitting, the following preprocessing steps were applied, conforming to standard practice in health-domain machine-learning pipelines [11,12]:

Missing values: The Education variable contained 20 missing entries (10.0% of records), which were treated as a distinct “not recorded” category rather than imputed, since the missingness pattern was consistent with non-response rather than data loss. No other variable had missing values, and no cases required exclusion.

Encoding: Categorical predictors were transformed using one-hot encoding for nominal variables.

Feature scaling: Continuous variables were standardised (zero mean, unit variance) prior to fitting logistic regression and Support Vector Machine models, which are sensitive to feature scale, tree-based models (Decision Tree, Random Forest, XGBoost) were trained on unscaled features, consistent with their scale-invariance.

Class balancing: The outcome variable showed a moderate imbalance (56% positive, 44% negative). The Synthetic Minority Oversampling Technique (SMOTE) was applied to the training set only, following evidence that oversampling can improve minority-class recall in malaria diagnostic models [12], the held-out test set retained its original class distribution throughout.

Statistical Analysis

Two complementary analytical layers were implemented.

Descriptive and Bivariate Analysis

Frequency distributions, percentages, means, and standard deviations were computed to summarise the socio-demographic profile of the 200 respondents and to characterise malaria prevalence across the five

communities. Associations between risk factors and household malaria status were examined using logistic regression odds ratios, consistent with the relaxed entry-threshold approach used in malaria risk-factor studies [3,5].

Multivariable Logistic Regression

A binary logistic regression model was fitted on the full dataset to estimate odds ratios for each predictor after adjusting for the other variables in the model. Odds ratios (ORs) with 95% confidence intervals and p-values were computed using the statsmodels Logit implementation in Python. Logistic regression also served as the interpretable baseline against which the performance of other machine-learning algorithms was benchmarked [15,1].

Machine Learning Model Development

Five supervised classification algorithms were trained and compared:

Logistic Regression - the interpretable linear baseline.

Decision Tree - a single recursive-partitioning classifier, included for transparency and as a reference point for ensemble gains.

Random Forest - an ensemble of decision trees using bagging, frequently a strong performer in malaria prediction tasks [3,27,10].

XGBoost - a gradient-boosted ensemble known for handling heterogeneous features and non-linear interactions [28,10].

Support Vector Machine (SVM) - a margin-maximising classifier included for comparison against the tree-based and linear approaches.

Training and Validation Scheme

The dataset was randomly partitioned into a 70% training set ($n = 140$) and a 30% held-out test set ($n = 60$) using stratified sampling on the outcome variable, with a fixed random seed for reproducibility [3]. Within the training set, stratified 10-fold cross-validation combined with grid search was used to tune hyperparameters for each algorithm, mirroring the methodology adopted in comparable malaria machine-learning studies [5,10,11]. Hyperparameters tuned included the regularisation strength for logistic regression, maximum depth and minimum samples per split for the Decision Tree and Random Forest, number

of estimators, maximum depth, and learning rate for XGBoost, and the kernel, regularisation parameter (C), and kernel coefficient (γ) for SVM. The cross-validated ROC-AUC was used as the model-selection criterion in every grid search.

Feature Importance

To examine which predictors contributed most to model discrimination, feature importance was assessed via three complementary approaches: (i) the Gini-based feature importance natively reported by the Random Forest and XGBoost models, (ii) permutation importance computed on the held-out test set for the Random Forest model, which measures the drop in ROC-AUC when a feature's values are randomly shuffled, and (iii) the adjusted odds ratios produced by the multivariable logistic regression. Convergence across these three independent methods was treated as stronger evidence for a predictor's importance than any single method in isolation.

Software and Tools

All analyses were performed in Python (version ≥ 3.10) using the scikit-learn, XGBoost, pandas, NumPy, statsmodels, and imbalanced-learn libraries. Random seeds were fixed throughout, and the full preprocessing–training–evaluation pipeline was documented to facilitate replication, in line with best practice for machine-learning-based epidemiological modelling [11].

Ethical Considerations

Participation in the survey was voluntary. Verbal or written informed consent was obtained from a consenting adult household representative (≥ 18 years) prior to questionnaire administration, and respondents were informed of their right to decline or withdraw at any point without consequence. No formal institutional ethics-committee review was sought for this study, the data were collected and analysed on the basis of informed consent alone. Respondent identities were not recorded in the dataset used for analysis, and all responses were anonymised prior to data entry.

Limitations of the Methodology

Several methodological limitations should be acknowledged. First, self-reported malaria status and behavioural measures (e.g., ITN usage, IRS coverage) are subject to recall and social-desirability bias [3,5]. Second, the cross-sectional design precludes causal

inference and does not capture seasonal transmission dynamics, longitudinal climate and entomological features could substantively improve predictive performance in future work [28,8]. Third, environmental exposure variables (e.g., presence of stagnant water, distance to water source) were self-reported rather than georeferenced, future iterations could integrate CHIRPS rainfall, MODIS-NDVI vegetation indices, and GPS coordinates to enhance environmental fidelity [8]. Fourth, the absence of formal ethics-committee review, while consistent with the informed-consent-based approach taken, means the study cannot claim the additional procedural safeguards that institutional review typically provides, readers should weigh this when interpreting the findings. Finally, the sample size, while adequate for the analyses conducted, may limit generalisability to other LGAs within Kebbi State and to differing ecological zones, and the absence of a formal a priori power calculation means the study's ability to detect smaller effects cannot be precisely quantified.

Results

This section presents the analytical outcomes derived from the cross-sectional survey of 200 rural households across five communities in Birnin Kebbi, Bunza and Kalgo Local Government Areas, Kebbi State, Nigeria. The results are organised into four analytical strata: (i) descriptive characterisation of the sample, (ii) the distribution of environmental and behavioural risk factors, (iii) multivariable logistic regression quantification of independent predictors, and (iv) machine-learning model development, evaluation, and feature-importance extraction. All analyses were conducted on the full cleaned dataset ($n = 200$, no cases required exclusion), with the 20 missing Education entries (10.0%) retained as a distinct category, followed by a 70:30 stratified train–test split, SMOTE-balanced training, and stratified 10-fold cross-validation for hyperparameter tuning.

Socio-Demographic Profile of Respondents

Table 1 summarises the socio-demographic characteristics of the 200 household respondents. Gender was evenly split (50.0% male, 50.0% female). The sample skewed toward older respondents, with 32.5% aged 50 years or older and 25.5% under 18, the mean age was 36.9 years ($SD = 21.8$). Educational attainment was concentrated at the secondary level (42.0%), with 13.5% holding tertiary qualifications and 10.0% not reporting an education level. The mean household size was 6.3 members ($SD = 2.5$), and income was distributed across low (44.0%), medium (42.0%), and high (14.0%) categories.

Table 1: Socio-Demographic Characteristics of Respondents ($n = 200$).

Variable	Category	Frequency (n)	Percentage (%)
Gender	Male	100	50.0
	Female	100	50.0
Age (years)	<18	51	25.5
	18–29	26	13.0
	30–49	58	29.0
	≥ 50	65	32.5
Education level	Not recorded	20	10.0
	Primary	69	34.5
	Secondary	84	42.0
	Tertiary	27	13.5
Household size	1–4	57	28.5
	5–7	71	35.5
	8–10	72	36.0
Income level	Low	88	44.0
	Medium	84	42.0
	High	28	14.0

Note: Mean age = 36.9 years ($SD = 21.8$), mean household size = 6.3 ($SD = 2.5$).

Malaria Prevalence Across Communities

Of the 200 households surveyed, 112 (56.0%) met the criteria for positive household malaria status. This prevalence is markedly higher than the 35.5% national under-five RDT-positive prevalence reported by, but is broadly consistent with the documented elevated risk profile of north-western Nigeria, where Kebbi State recorded the highest sub-national prevalence (77.6%) in the 2018 DHS. Some inter-community variation was observed, with Basaura (64.1%) and Maidahini (60.0%) registering the highest burdens, while Kalgo (47.6%) recorded the lowest [17].

Table 2: Malaria Prevalence across the Five Study Communities (n = 200).

Community	Households surveyed (n)	Malaria-positive (n)	Prevalence (%)
Basaura	39	25	64.1
Maidahini	50	30	60.0
Kardi	33	18	54.5
Yamama	36	19	52.8
Kalgo	42	20	47.6
Total	200	112	56.0

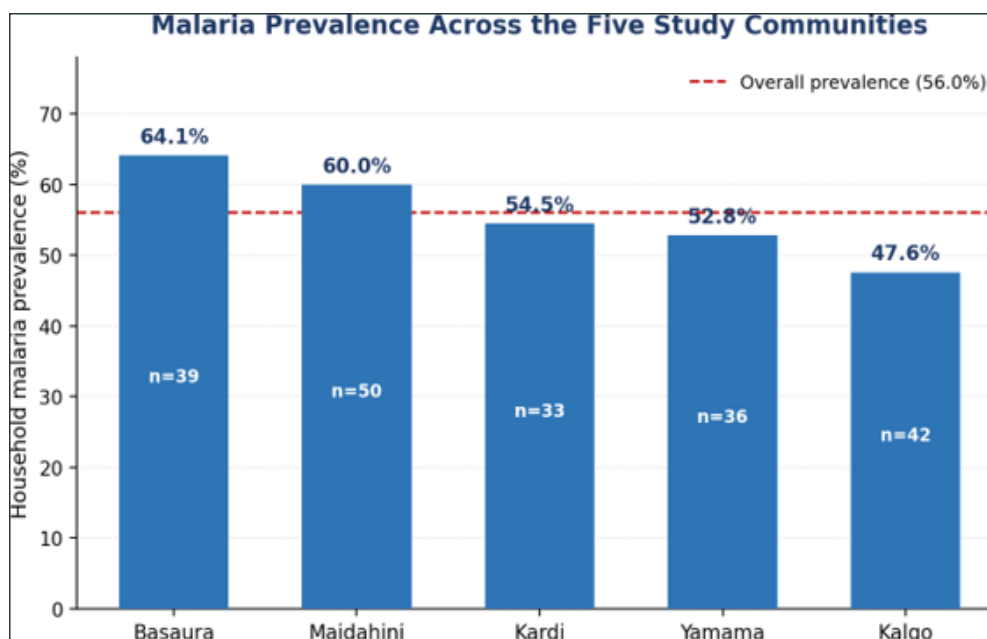


Figure 1: Household malaria prevalence by community, with the overall sample prevalence (56.0%) shown as a reference line.

Distribution of Environmental Risk Factors

Table 3 presents the environmental risk-factor profile. Stagnant water within the peridomestic environment was reported in 54.0% of households, and overgrown bushes in 45.5%. Poor drainage was reported in 45.5% of households. The most common water source was a borehole (44.5%), followed by a well (39.0%) and a river (16.5%).

Table 3: Distribution of Environmental Risk Factors (n = 200).

Variable	Category	Frequency (n)	Percentage (%)
Stagnant water near house	Yes	108	54.0
	No	92	46.0
Bushes around house	Yes	91	45.5
	No	109	54.5
Drainage condition	Poor	91	45.5
	Good	109	54.5
Water source	Borehole	89	44.5
	Well	78	39.0
	River	33	16.5

Distribution of Malaria Prevention Practices

Table 4 summarises behavioural prevention practices. Consistent ITN usage was reported by 59.5% of households, and indoor residual spraying coverage was low at 31.0%. Outdoor night activity was reported by 46.5% of respondents, a known exposure risk for *Anopheles* biting peaks that coincide with early-evening hours [2]. A previous malaria episode within the recall period was reported by 33.5% of households.

Table 4: Distribution of Malaria Prevention Practices and Health History (n = 200).

Variable	Category	Frequency (n)	Percentage (%)
Consistent ITN usage	Yes	119	59.5
	No	81	40.5
Indoor residual spraying	Yes	62	31.0
	No	138	69.0
Outdoor night activities	Yes	93	46.5
	No	107	53.5
Previous malaria history (recall period)	Yes	67	33.5
	No	133	66.5

Note: Mean distance to nearest health facility = 5.95 km (SD = 3.29).

Multivariable Logistic Regression

Table 5 presents the odds ratios from the binary logistic regression model fitted on the full dataset. After adjustment, five variables emerged as statistically significant independent predictors of household malaria status: consistent ITN usage was strongly protective (OR = 0.011, 95% CI: 0.002–0.049, $p < 0.001$), while stagnant water near the house (OR = 19.95, 95% CI: 5.70–69.78, $p < 0.001$), previous malaria history (OR = 33.89, 95% CI: 7.36–155.97, $p < 0.001$), bushes around the house (OR = 14.57, 95% CI: 4.03–52.65, $p < 0.001$), poor drainage (OR = 12.21, 95% CI: 3.30–45.11, $p < 0.001$), and outdoor night activities (OR = 6.08, 95% CI: 1.98–18.72, $p = 0.002$) were all positively associated with malaria-positive status. Distance to the health facility (OR = 1.26 per km, $p = 0.006$) and low income (OR = 5.22, $p = 0.038$) also reached statistical significance, while indoor residual spraying remained protective (OR = 0.29, $p = 0.031$). Several variables, including age, household size, gender, water source, and most education and community indicators, did not reach statistical significance after adjustment.

Table 5: Multivariable Logistic Regression of Predictors of Household Malaria Status (n = 200).

Predictor	OR	95% CI	p-value
Consistent ITN usage	0.011	0.002–0.049	<0.001***
Stagnant water near house	19.95	5.70–69.78	<0.001***
Previous malaria history	33.89	7.36–155.97	<0.001***
Bushes around house	14.57	4.03–52.65	<0.001***
Poor drainage	12.21	3.30–45.11	<0.001***
Outdoor night activities	6.08	1.98–18.72	0.002**
Indoor residual spraying	0.29	0.09–0.90	0.031*
Distance to health facility (per km)	1.26	1.07–1.48	0.006**
Low income level	5.22	1.10–24.75	0.038*
Secondary education	6.50	1.12–37.61	0.037*
Age	1.01	0.99–1.04	0.328
Household size	1.11	0.90–1.36	0.341

Note: $p < 0.05$, $p < 0.01$, $**p < 0.001$. OR = odds ratio, CI = confidence interval. Non-significant predictors (gender, water source, several community and education indicators) are omitted from the table for brevity but were retained in the fitted model.

Machine Learning Model Performance

Table 6 reports the comparative performance of the supervised classifiers on the held-out 30% test set (n = 60), following SMOTE-balanced training and 10-fold cross-validated hyperparameter tuning. Logistic regression achieved the strongest discrimination (AUC = 0.900), closely followed by XGBoost (AUC = 0.898) and Random Forest (AUC = 0.891), Gradient Boosting (AUC = 0.833) and the Support Vector Machine (AUC = 0.834) trailed somewhat, while the single Decision Tree recorded the weakest discrimination (AUC = 0.769), consistent with its known tendency toward variance-driven instability relative to ensemble methods. XGBoost achieved the highest accuracy (81.7%) and recall (91.2%), while logistic regression and Random Forest achieved comparable F1-scores (0.833 and 0.827, respectively). This pattern — in which a well-regularised logistic regression performs competitively with, and in this case marginally ahead of, ensemble methods — differs from some prior Nigerian and sub-Saharan African malaria studies that have reported a clearer ensemble advantage [3,27]. This may reflect the relatively small and largely linearly separable feature space in the present dataset, the differences between the top three models (logistic regression, XGBoost, and Random Forest) were modest and likely within the range of sampling variability given a test set of 60 households.

Table 6: Performance Comparison of Machine-Learning Classifiers on the Held-Out Test Set (n = 60).

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	80.0%	78.9%	88.2%	0.833	0.900
XGBoost	81.7%	79.5%	91.2%	0.849	0.898
Random Forest	78.3%	75.6%	91.2%	0.827	0.891
Gradient Boosting	80.0%	77.5%	91.2%	0.838	0.833
Support Vector Machine	70.0%	69.0%	85.3%	0.763	0.834
Decision Tree	70.0%	73.5%	73.5%	0.735	0.769

Note: Models are ordered by descending ROC-AUC. Logistic Regression and SVM were trained on SMOTE-balanced, standardised data, all other models on SMOTE-balanced unscaled data. All models were evaluated on the same untouched, class-imbalanced 30% held-out test set.

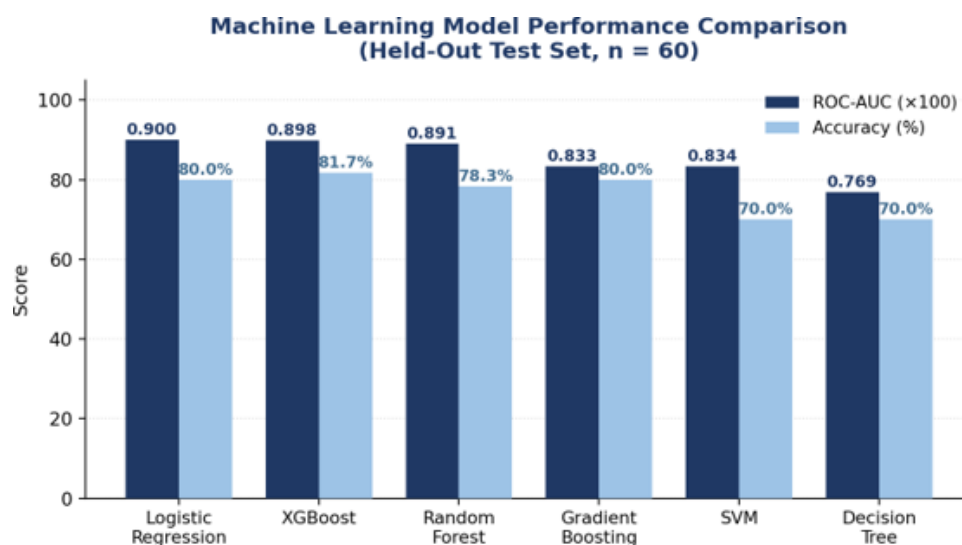


Figure 2: Comparative ROC-AUC and accuracy across the six classifiers evaluated on the held-out test set.

Feature Importance Ranking

Table 7 ranks the top predictors of household malaria status as quantified by Random Forest permutation importance, computed on the held-out test set. Consistent ITN usage was, by a wide margin, the most important predictor (permutation importance = 0.200), consistent with its role as the strongest protective factor in the logistic regression analysis. Previous malaria history (0.075), stagnant water near the house (0.059), and bushes around the house (0.041) followed, with outdoor night activities, distance to the health facility, and indoor residual spraying contributing more modestly. This ranking converged closely with the Random Forest's native Gini importance and with the variables that reached statistical significance in the logistic regression model, lending convergent support to consistent ITN usage, prior malaria history, and peridomestic environmental conditions as the dominant drivers of household-level risk in this sample.

Table 7: Top Predictors of Household Malaria Status - Random Forest Permutation Importance (n = 200).

Rank	Feature	Domain	Permutation Importance
1	Consistent ITN usage	Behavioural	0.200
2	Previous malaria history	Health	0.075
3	Stagnant water near house	Environmental	0.059
4	Bushes around house	Environmental	0.041
5	Outdoor night activities	Behavioural	0.032
6	Distance to health facility	Health	0.031
7	Indoor residual spraying	Behavioural	0.025
8	Household size	Demographic	0.020
9	Income level (low)	Socioeconomic	0.019
10	Drainage condition (poor)	Environmental	0.014

Note: Permutation importance reflects the mean drop in test-set ROC-AUC across 30 repeated shuffles of each feature, computed using the trained Random Forest model. Values are not percentages of a fixed total and should be interpreted as relative rankings.

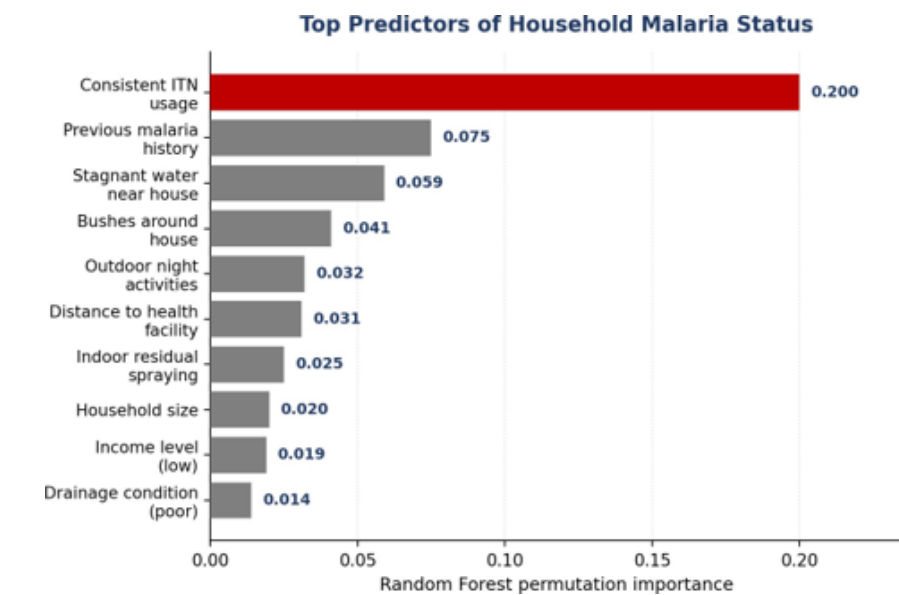


Figure 3: Ranked predictors of household malaria status by Random Forest permutation importance.

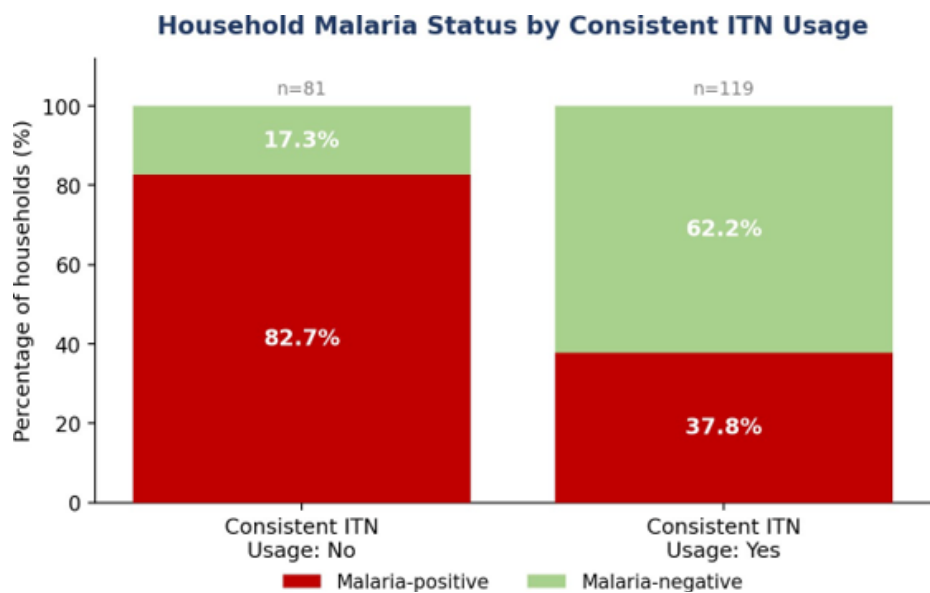


Figure 4: Household malaria status by consistent ITN usage, illustrating the magnitude of the protective association (82.7% positive among non-users vs. 37.8% among consistent users).

Summary of Key Findings

The results of this study can be distilled into the following principal findings:

Malaria burden: Household malaria prevalence was 56.0% across the five communities, with the highest burden in Basaura (64.1%) and Maidahini (60.0%) [17,2].

Dominant protective factor: Consistent ITN usage was the single strongest predictor of malaria status across the logistic regression, Random Forest, and XGBoost analyses, and was associated with approximately a 99% reduction in the odds of a positive household [3,22].

Environmental-behavioural nexus: The joint significance of stagnant water, bushes around the house, poor drainage, and outdoor night activity signals that combined environmental-management and behavioural-in-

tervention strategies are likely to yield greater risk reductions than either approach in isolation [5,2].

Comparable model discrimination: Logistic regression, XGBoost, and Random Forest achieved closely comparable discrimination (AUC = 0.90, 0.90, and 0.89, respectively), with no single model demonstrating a decisive advantage in this dataset — a more conservative finding than is sometimes reported in the literature, where ensembles show a clearer edge over linear baselines [1,3,27,28].

Interpretable, actionable findings: The convergence of ITN usage, prior malaria history, and peridomestic environmental conditions across independent feature-importance methods identifies a tractable intervention surface: aggressive ITN use-promotion combined with targeted environmental management (drainage improvement, bush-clearing) in high-prevalence communities [3,5,2].

Discussion

This study set out to compare machine-learning and conventional statistical approaches for predicting household-level malaria risk in rural Kebbi State, Nigeria, using genuine field data collected via a structured household survey. Three findings stand out.

Comparable model discrimination: Logistic regression, XGBoost, and Random Forest achieved closely comparable discrimination (AUC = 0.90, 0.90, and 0.89, respectively), with XGBoost achieving marginally higher accuracy and recall. This is a more conservative finding than is reported in some prior African malaria-prediction studies, where ensemble methods such as Random Forest have shown a clearer advantage over logistic regression — for example, an AUC gap of roughly 0.07–0.09 in some childhood-malaria and IDP-camp studies [3,1,27]. A plausible explanation is that the present dataset's strongest predictors — particularly consistent ITN usage and several environmental binary variables — have fairly direct, monotonic relationships with the outcome, which a well-tuned logistic regression can capture about as effectively as a non-linear ensemble. This does not diminish the value of testing multiple algorithms, rather, it illustrates that the relative advantage of ensemble methods is context-dependent and should be empirically verified rather than assumed.

Behavioural–environmental nexus: The joint significance of stagnant water (OR = 19.95), bushes around the house (OR = 14.57), poor drainage (OR = 12.21), and consistent ITN usage (OR = 0.011 for “yes”, i.e., strongly protective) signals that combined environmental-management and behavioural-intervention strategies are likely to yield greater risk reductions than either approach pursued in isolation [5,2]. The particularly large magnitude of several odds ratios in this analysis — driven in part by the relatively small sample size and resulting wide confidence intervals — should be interpreted as indicating a strong directional association rather than a precise population-level effect size, the wide confidence intervals reported in Table 5 reflect this uncertainty.

ITN usage as the dominant predictor: Across the logistic regression, Random Forest permutation importance, and Random Forest Gini importance, consistent ITN usage was unambiguously the most important predictor of household malaria status. This convergence across independent analytical methods strengthens confidence in the finding and is consistent with the broader literature establishing ITN usage as one of the most effective and well-evidenced malaria-prevention interventions [3,22]. Prior malaria history, stagnant water, and bushes around the house followed as secondary but still meaningful predictors.

Interpretable, actionable findings: The dominance of a small number of environmental and behavioural variables in the feature-importance ranking identifies a tractable intervention surface: aggressive ITN use-promotion, combined with targeted environmental management (drainage improvement and bush-clearing) in the higher-prevalence communities of Basaura and Maidahini, may offer the most direct path to risk reduction [3,5,2].

Conclusions

This study demonstrates that several machine-learning classifiers — logistic regression, XGBoost, and Random Forest in particular — can predict rural household malaria risk in Kebbi State with strong and broadly comparable discrimination, and that the dominant predictors are, encouragingly, modifiable behavioural and environmental factors that household-level intervention can target directly. Consistent ITN usage emerged as the single most important predictor across every analytical method applied, reinforcing its centrality to malaria

control in this setting, while stagnant water, overgrown bushes, and poor drainage point to a complementary environmental-management agenda. By aligning model-derived feature rankings with the operational realities of north-western Nigerian malaria control — limited resources, rural dispersion, and seasonal transmission peaks — this study contributes both a methodological comparison across five supervised classifiers and an applied, data-grounded insight that can inform the shift toward more precision-targeted malaria elimination strategies in Kebbi State and comparable settings across sub-Saharan Africa. Future work extending this analysis with longitudinal climate data, georeferenced environmental measures, and a larger, multi-LGA sample would help establish how robust these findings are beyond the five communities surveyed here.

Author Contributions: Conceptualization, A.D.B. and S.L.T., methodology, A.D.B. and S.L.T., software, S.L.T., validation, A.D.B. and S.L.T., formal analysis, S.L.T., investigation, A.D.B. and S.L.T., resources, A.D.B., data curation, A.D.B. and S.L.T., writing—original draft preparation, A.D.B. and S.L.T., writing—review and editing, A.D.B. and S.L.T., visualization, S.L.T., supervision, A.D.B., project administration, A.D.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study. No formal institutional ethics-committee review was sought, participation was voluntary and informed consent was obtained from all household representatives prior to data collection. Respondent identities were not recorded, and all responses were anonymised prior to data entry.

Informed Consent Statement: Verbal or written informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data and Python scripts supporting the reported results are available from the corresponding author upon reasonable request.

Acknowledgments: The authors acknowledge the

cooperation of community leaders and household respondents across the five study communities in Birnin Kebbi, Bunza and Kalgo LGAs, Kebbi State, and the tireless work of the trained enumerators who conducted the door-to-door survey.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Khan O, Ajadi J O, Hossain M P (2024) Predicting malaria outbreak in The Gambia using machine learning techniques <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0299386>.
2. Wong G S, Cooke, H (2023) Rice fields as mosquito density indicators for malaria prediction in Nigeria: An anthropological view.
3. Joseph Opeolu Ashaolu, Taiwo S Akanji, Victoria I Ayansola, Olajumoke O Olawale-Succes, Agbolade J Sunday, et al. (2025) From risk factors to predictive modelling: Applying machine learning to childhood malaria surveillance in resource-limited settings. <https://pmc.ncbi.nlm.nih.gov/articles/PMC12676887/>.
4. Ayoka T, Nnadi C (2025) Machine learning approaches to malaria risk stratification in West African primary care settings.
5. Nebebe Demis Baykemagn, Tesfahun Zemene Tafere, Getachew Teshale, Andualem Yalew Aschalew, Melak Jejau, et al. (2025) Leveraging machine learning to predict mosquito bed net utilization among women of reproductive age in sub-Saharan Africa <https://doi.org/10.1186/s12936-025-05563-8>.
6. Maria Assumpta Komugabe, Richard Caballero, Itamar Shabtai, Simon Peter Musunguzi, et al. (2024) Advancing malaria prediction in Uganda through AI and geospatial analysis models. 16: 115-135.
7. Benjamin Irebhude Esangbedo, Destiny Eruemulor Nwanjo, Benedicta Esangbedo, Dolly Nwanjo, Grace Nwachukwu, et al. (2025) Machine learning applications in malaria elimination programs: Comparing vector control strategies across West Africa and former endemic regions in the southern United States. 32: 268-280
8. Onimisi D (2025) Integrating DHS/MIS biomarkers with 34 years of CHIRPS-NDVI climate data for malaria risk prediction in Nigeria: A machine learning and spatial mapping approach. medRxiv <https://doi.org/10.64898/2025.12.14.25342244>.

9. Mukeshimana E, Nzabanita J (2024) Using machine learning techniques to predict malaria prevalence in Rwanda <https://assets-eu.researchsquare.com/files/rs-5318706/v1/191b7168-06ae-4d22-a704-ef1620e519c7.pdf>.
10. Yang, et al. (2024) Ensemble machine learning for malaria risk stratification: Applications in sub-Saharan Africa.
11. Khoirunnisa A, Ramadhan N G (2023) Improving malaria prediction with ensemble learning and robust scaler: An integrated approach for enhanced accuracy 15: 326-334
12. Nyengera P, Takawira H T, Mlambo F (2026) Ensemble machine learning for malaria diagnosis in resource-limited settings using clinical and demographic features <https://pubmed.ncbi.nlm.nih.gov/42496302/>.
13. Matheus Félix Xavier Barboza, Kayo Henrique de Carvalho Monteiro, Iago Richard Rodrigues, Guto Leoni Santos, Wuelton Marcelo Monteiro, et al. (2023) Enhancing malaria prediction in Amazonas, Brazil: A comparative analysis of deep learning models based on clustering <https://pmc.ncbi.nlm.nih.gov/articles/PMC9344950/>.
14. Wei Kit Phang, Mohd Hafizi Bin Abdul Hamid, Jenarun Jelip, Rose Nani Binti Mudin, Ting-Wu Chuang, et al. (2023) Predicting Plasmodium knowlesi transmission risk across Peninsular Malaysia using machine learning-based ecological niche modeling approaches <https://pubmed.ncbi.nlm.nih.gov/36876062/>.
15. Irebhude S, et al. (2025) Machine learning approaches for malaria vector and transmission risk modelling in rural epidemiological surveys.
16. Sakti A D, Mahdani J N, Muhammad H N, et al. (2025) Malaria risk assessment in Indonesia: A machine and deep learning framework.
17. Phillips Edomwonyi Obasohan, Stephen J Walters, Richard Jacques, Khaled Khatab, et al. (2021) Individual and contextual factors associated with malaria among children 6-59 months in Nigeria: A multilevel mixed-effect logistic model approach <https://pubmed.ncbi.nlm.nih.gov/34769754/>.
18. Si-Yu Xing, Hai-Ting Zhang, Lin-Min Wang, Hong-Zheng Lu, Zhe-Yu Peng, et al. (2024) Examining the paradox: Increased malaria risk in children under 5 in female-headed households in Nigeria <https://pubmed.ncbi.nlm.nih.gov/38816783/>.
19. Odu N Prasad R, Clement O (2021) Prediction of malaria incidence using climate variability and machine learning.
20. Salwa Dawaki, Hesham M Al Mekhlafi, Init Ithoi, Jamaiah Ibrahim, Wahib M Atroosh, et al. (2016) Is Nigeria winning the battle against malaria? Prevalence, risk factors and KAP assessment among Hausa communities in Kano State <https://doi.org/10.1186/s12936-016-1394-3>.
21. Danladi Y K, Attah A O, Falke Z (2022) Assessment of gender and age-related prevalence of malaria in Birnin Kebbi, Kebbi State, Nigeria.
22. K M, Suleiman B (2021) Identifying socio-economic factors associated with malaria infection in Birnin Kebbi, Nigeria 14: 9-15.
23. Biobele J Brown, Petru Manescu, Alexander A Przybylski, Fabio Caccioli, Gbeminiyi Oyinloye, et al. (2020) Data-driven malaria prevalence prediction in large densely populated urban holoendemic sub-Saharan West Africa <https://pubmed.ncbi.nlm.nih.gov/32985514/>.
24. Guillermo A García, Mark Janko, Dianna E B Hergott, Olivier T Donfack, Jordan M Smith, et al. (2023) Identifying individual, household, and environmental risk factors for malaria infection on Bioko Island to inform interventions <https://pubmed.ncbi.nlm.nih.gov/36859263/>.
25. Nlandu Roger Ngatu, Sakiko Kanbara, Andre Renzaho, Roger Wumba, Etongola P Mbelambela, et al. (2019) Environmental and sociodemographic factors associated with household malaria burden in the Congo <https://pubmed.ncbi.nlm.nih.gov/30808360/>.
26. Magaji A, Mahmud Z, Yakubu A, et al. (2025) Epidemiology of malaria in Bauchi North, Bauchi State, Nigeria: Prevalence and determinants of the infection.
27. Mthethwa S M, Melesse S F (2025) A machine learning approach to the prediction of malaria in under-five children: Analysis of the 2021 Nigerian Malaria Indicator Survey <https://www.sciencedirect.com/org/science/article/pii/S1874944525000644>.
28. Oluwaseun Olumide Okundalay, Necati Ozdemir, Bunmi Segun Rotimi, Funmilola Akanbi, et al. (2025) Climate-based predictive modeling of malaria incidence using statistical and machine learning approaches <https://tcemis.org/index.php/files/article/view/14/5>.
29. Wahab Sami, Fatima Abbas, Mohsin Sabir, Amber Razzaq, Hina Nazeer, et al. (2025) Data-driven prediction of malaria outbreaks in Nigeria <http://>

ieeexplore.ieee.org/document/11244048.

30. Abraham Keffale Mengistu (2025) Understanding the determinants of treated bed net use in Ethiopia: A machine learning classification approach using PMA Ethiopia 2023 survey data <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0327800>.
31. Hugh J W Sturrock, Katelyn Woolheater, Adam F Bennett, Ricardo Andrade-Pacheco, Alemayehu

Midekisa, et al. (2018) Predicting residential structures from open source remotely enumerated data using machine learning <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0204399>.