Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predicting malaria incidence using meteorological, environmental, and spatial factors in Guinea-Bissau: a machine learning approach

Marcelino Correia NANQUE
cnanque.marcelino@univ-thies.sn
Université de Thiès
Mouhamad M. ALLAYA
Université de Thiès
Mor Absa LOUM
Université de Thiès
Vasco Na DUM
Université de Thiès
José E. NANTE
Université de Thiès
Lucien D. GNING
Université Gaston Berger
Mouhamed Ould HAMED
Programa Nacional de Luta contra Paludismo (PNLT)
Ousmane SY
Cheikh Anta Diop University
Almamy Y. LY
Université de Thiès
Vítor M.L.C. PEREIRA
Université de Thiès
Baltazar CA
Bandim Health Project
Mamadou BOUSSO
Université de Thiès
Jean Louis Abdourahim NDIAYE
Université de Thiès

Research Article

Keywords: Malaria prediction, Machine learning, Climate variables, XGBoost, Time-series models, Guinea-Bissau, Public Health

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10972587/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Predicting malaria incidence using meteorological, environmental, and spatial factors in Guinea-Bissau: a machine learning approach

Marcelino Correia NANQUE[*1,7], Mouhamad M. ALLAYA[2], Mor Absa LOUM[3], Vasco Na DUM[1,4], José E. NANTE[1,4], Mouhamed Ould HAMED[4], Ousmane SY[5], Lucien D. GNING[6], Almamy Y. LY[8], Vítor M.L.C. PEREIRA[8], Baltazar CA[9], Mamadou BOUSSO[2], Jean Louis Abdourahim NDIAYE[10]

*** Corresponding Author**
Marcelino Correia NANQUE
cnanque.marcelino@univ-thies.sn
sr.marcel86@gmail.com
https://orcid.org/0009-0003-9914-0854

Abstract

Background: Malaria remains a major public health problem in Guinea-Bissau, where transmission is strongly influenced by climatic and environmental conditions. Despite progress, its incidence exhibits significant seasonal and geographic variations. This study aimed to analyze the influence of meteorological factors on malaria incidence, evaluate the ability of machine learning models for trend prediction, and conduct a comparative evaluation against conventional time-series approaches across the health regions of Guinea-Bissau.

Methods: The study used 924 monthly observations collected across Guinea-Bissau's 11 health regions between January 2016 and December 2022, considering 28 candidate variables spanning geographical, demographic, epidemiological, environmental, and climatic dimensions. Three machine learning models Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM) were developed and benchmarked against conventional time-series models, namely SARIMA, ETS, and GLMM with Splines. Performance was evaluated using RMSE, MAE, and R^2, validated across temporal horizons up to 2022, and supported by residual diagnostics and spatial autocorrelation analyses (Moran's I). SHAP values were used to interpret variable importance.

Results: The XGBoost model achieved the best predictive performance, with an R^2 of 0.78 (95% CI: 0.52–0.97), alongside an RMSE of 2.27 (95% CI: 1.28-3.15) and an MAE of 1.74 (95% CI: 0.89–2.63) on the test set (2021), as well as strong validation metrics an R^2

of 0.91(95% CI: 0.82-0.98), an RMSE of 1.57 (95% CI: 0.87-2.24) and an MAE of 1.22 (95% CI: 0.65-1.85) on the 2022 horizon. Machine learning algorithms consistently outperformed conventional temporal models (SARIMA, ETS, and GLMM with Splines) in capturing non-linear relationships. The models maintained robust generalization without significant residual spatial autocorrelation.

Conclusions: Machine learning models, particularly XGBoost, significantly outperform conventional time-series approaches in predicting malaria dynamics in Guinea-Bissau. Integrating epidemiological, climatic, and environmental data improves the identification of high-risk periods, contributing to early warning systems and optimized control interventions.

Keywords: Malaria prediction, Machine learning, Climate variables, XGBoost, Time-series models, Guinea-Bissau, Public Health.

Introduction

According to the World Health Organization's (WHO) 2024 World Malaria Report, malaria remains one of the world's leading infectious diseases, with approximately 249 million cases and 608,000 deaths recorded globally in 2023. The disease burden is heavily concentrated in the WHO African Region, which accounts for nearly 95% of global cases and deaths, disproportionately affecting vulnerable groups such as children under five and pregnant women [1, 2].

Malaria transmission is driven by complex, non-linear interactions among environmental, climatic, and socioeconomic factors. Meteorological variables such as precipitation, temperature, and relative humidity directly regulate the survival, reproduction, and vectorial capacity of *Anopheles* mosquitoes, as well as the extrinsic incubation period of *Plasmodium* parasites. Given the intricate nature of these environmental-epidemiological relationships, traditional linear approaches often fall short, prompting the increasing adoption of machine learning and spatial modeling to enhance predictive accuracy and early warning capabilities [3].

In Guinea-Bissau, malaria represents a persistent public health crisis, characterized by endemic transmission and pronounced spatial-temporal variations [4–6]. The disease is primarily caused by *Plasmodium falciparum* and accounts for a substantial proportion of

outpatient visits, hospitalizations, and deaths, particularly during the peak transmission window toward the end of the rainy season between October and November.

The global trend in malaria incidence has also been observed in Guinea-Bissau over the past two decades, particularly among children under five. The disease is endemic throughout the country and is characterized by seasonal transmission that is strongly influenced by climatic conditions, including rainfall, temperature, and humidity [7–9]. These environmental conditions promote the proliferation of Anopheles mosquitoes, which transmit the Plasmodium falciparum parasite, the most virulent species in sub-Saharan Africa [10, 11].

National surveillance data from the National Malaria Control Program (NMCP) highlight that heavy seasonal rainfall drives the proliferation of aquatic larval habitats, triggering sharp surges in vector density and clinical incidence [12].

Spanning a tropical climate with distinct wet and dry seasons, Guinea-Bissau exhibits distinct regional vulnerabilities. Despite ongoing control interventions such as the distribution of insecticide-treated nets, rapid diagnostic tests, and artemisinin-based combination therapies annual fluctuations in incidence persist. These variations indicate that localized climatic and environmental drivers significantly modulate transmission dynamics [13].

Understanding and quantifying the precise relationships between meteorological factors and malaria incidence in Guinea-Bissau is therefore critical [14]. Addressing this knowledge gap will inform the refinement of predictive early warning systems and optimize the spatial-temporal targeting of malaria control strategies across the country.

Methods

Study area

Guinea-Bissau is a West African country, bordered to the west by the Atlantic Ocean and located between the Republic of Senegal to the north and the Republic of Guinea to the south and east (12°N, 15°W).

The country covers an area of approximately 36,125 km² and has a population of around 2,046,289 inhabitants according to recent demographic estimates, According to the Harmonized Household Living Conditions Survey conducted in Guinea-Bissau, important

socioeconomic disparities persist across regions [15]. Administratively and in terms of public health, Guinea-Bissau is divided into 11 health regions and 117 health zones, and includes an archipelago of 88 islands, 21 of which are inhabited, mainly located in the Bijagós Archipelago [5]. From an ecological point of view, four major ecosystems have been identified: The Guinean savannah, which predominates in the Gabú region in the east of the country; Forest areas in the regions of Quinara and Tombali; Coastal areas and mangroves in the regions of Cacheu, Oio and Biombo; The densely populated urban area of the Bissau Autonomous Sector (SAB) (Fig.1). These different ecosystems strongly influence the dynamics of malaria transmission, in particular thanks to the availability of breeding sites for Anopheles mosquitoes [6].

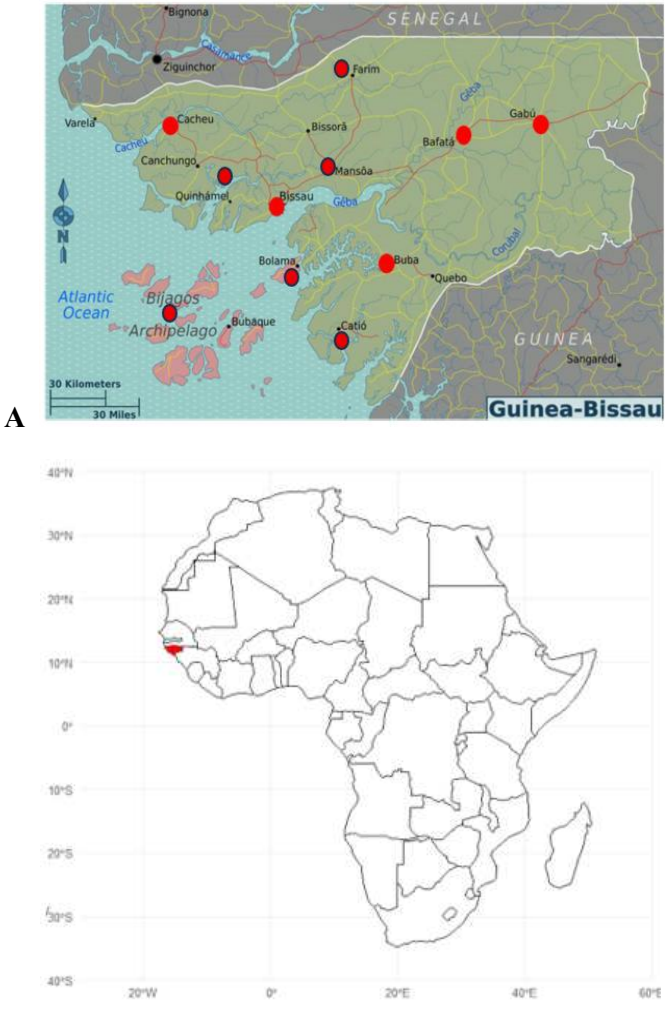


Fig. 1. Geographic location of the study area. (**A**) Administrative map of Guinea-Bissau showing the health regions included in the analysis of malaria incidence between 2016 and 2022. (**B**) Location of Guinea-Bissau in West Africa.

Guinea-Bissau enjoys a tropical climate of the Sudano-Guinean type, characterized by two main seasons: A dry season, from November to May, dominated by harmattan and characterized by low humidity and practically no rainfall; A rainy season, from June to October, marked by heavy rainfall and high relative humidity. Annual precipitation typically ranges from 1,200 mm to 2,500 mm, with a marked spatial gradient between coastal regions and the interior of the country [16]. The southern and coastal regions, especially Tombali, Quinara, and Cacheu, receive the most rainfall, while eastern regions such as Gabú and Bafatá have relatively low rainfall levels.

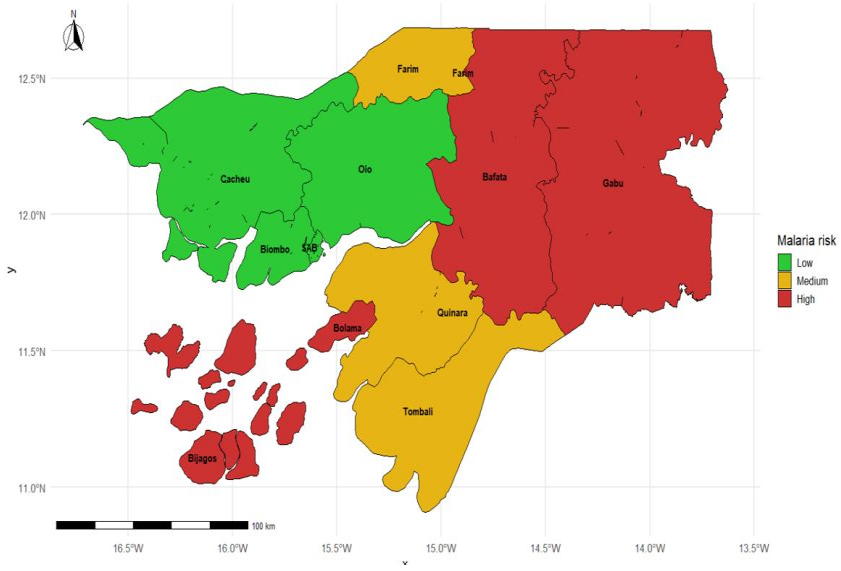


Fig. 2. Classification of regions according to the level of risk of malaria in Guinea-Bissau, NMCP 2023

Malaria transmission in Guinea-Bissau is endemic and stable, with higher transmission rates during and immediately after the rainy season, when environmental conditions favor the proliferation of vectors. The disease is one of the leading causes of morbidity and patient visits to health facilities in the country (Fig. 2). Several studies have shown that climate variability, particularly interannual fluctuations in precipitation, temperature, and relative humidity, associated with the dynamics of malaria transmission in West Africa, including Guinea-Bissau [13].

Data Sources and Processing

This study used epidemiological, climatic, environmental, and demographic data covering the period from January 2016 to December 2022 in Guinea-Bissau. Epidemiological data were obtained from the

national health surveillance system and included confirmed cases of malaria diagnosed by rapid diagnostic test (RDT) or microscopy, in accordance with World Health Organization recommendations [1]. Meteorological variables such as air temperature, precipitation, relative humidity, wind speed, and solar radiation were obtained from international remote sensing and climate reanalysis platforms (including NASA and global climate databases) and aggregated to a monthly resolution [17]. Environmental variables included the Normalized Difference Vegetation Index (NDVI) derived from MODIS, soil moisture from the Global Land Data Assimilation System (GLDAS), elevation from the Shuttle Radar Topography Mission (SRTM), and regional boundaries from GADM. Prior to modeling, data underwent a quality control procedure. Variables with more than 20% missing data were excluded. For minor gaps, missing values were imputed using linear temporal interpolation; for longer periods, regional averages from the same region were applied [14, 18, 19]. Continuous variables were standardized to improve model convergence, and the dataset was divided chronologically into training (2016–2020), test (2021), and validation (2022) sets to prevent information leakage [20, 21].

To evaluate potential biases related to healthcare utilization during the COVID-19 pandemic (2020–2022) and test model stability, sensitivity analyses were conducted by training models exclusively on pre-pandemic data (2016–2019) and evaluating them across the pandemic period [22, 23].

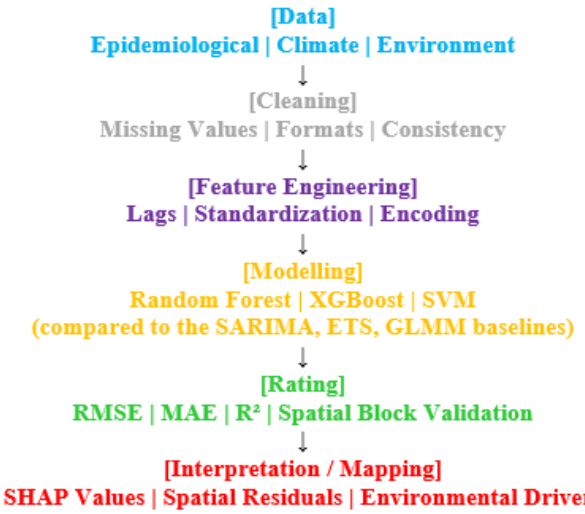


Fig. 3. Methodological workflow pipeline illustrating the sequential steps from data ingestion and cleaning, feature engineering, machine learning modeling (benchmarked against baseline models), performance rating via spatial block validation, to final model interpretation and mapping.

To ensure a systematic and reproducible analytical process, the methodological framework was structured into a multi-step workflow pipeline (Fig. 3). The pipeline encompasses initial data ingestion from epidemiological, climatic, and environmental sources, followed by rigorous data cleaning and quality control. Subsequently, feature engineering procedures including standardization and temporal alignment prepared the dataset for predictive modeling. Machine learning algorithms (Random Forest, XGBoost, and Support Vector Machines) along with conventional baseline models (SARIMA, ETS, and GLMM) were trained and evaluated using robust metrics within a spatial block cross-validation framework. Finally, model interpretation was achieved through SHAP values, spatial residual analysis, and environmental driver identification [24, 25].

Variable Construction and Selection

An initial set of 28 candidate variables spanning geographical, demographic, epidemiological, environmental, and climatic dimensions was considered (Supplementary Table S1). To capture vector-borne transmission dynamics, lagged variables for precipitation (from CHIRPS) and NDVI covering temporal windows of 1 to 3 months were initially generated [17]. However, sensitivity analyses regarding temporal structures demonstrated that these intermediate lags introduced high collinearity and redundancy with contemporaneous metrics; consequently, the final modeling framework retained contemporaneous environmental variables alongside specific selected factors (such as soil moisture).

An iterative feature selection process was then implemented based on the gain importance of machine learning algorithms (Random Forest and XGBoost) and the rigorous mitigation of multicollinearity assessed via Pearson's correlation matrix is presented in Supplementary Figure S1 and Variance Inflation Factor (VIF) tests [26]. During this filtering process: Precipitation and NDVI lags were systematically excluded due to informational redundancy, given that current-month variables optimally captured the predictive signal.

Relative humidity was excluded because of its strong correlation with soil moisture and temperature-related variables [27, 28]. Seasonal malaria chemoprevention (zone_qps) was explicitly excluded to prevent endogeneity bias and reverse causality, as public health interventions are proactively targeted toward areas with higher endemic risk. The final predictive model retained only predictors that provided complementary epidemiological and environmental information while minimizing redundancy, multicollinearity, and potential sources of bias. Model performance across these configurations was evaluated using the root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2) [29].

Machine Learning Models and Validation Framework

Three supervised machine learning algorithms were implemented to model and predict malaria incidence in Guinea-Bissau: Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM). These complementary algorithms were selected to handle complex non-linear interactions among explanatory variables: Random Forest was introduced by Leo Breiman in 2001 [30], it builds an ensemble of decision trees from bootstrap samples to reduce overfitting; XGBoost developed by Tianqi Chen and Guestrin in 2016, sequentially constructs trees via gradient boosting and regularization to optimize predictive performance [31]; and SVM utilizes kernel functions (specifically the Radial Basis Function - RBF) to map and separate complex non-linear relationships in high-dimensional feature spaces. Is a supervised learning algorithm developed by Corinna CORTES and Vladimir VAPNIK in 1995. The Support Vector Machine model is a supervised learning algorithm developed by Corinna CORTES and Vladimir VAPNIK in 1995. It identifies an optimal hyperplane that maximizes the separation between observations in feature space. SVM is particularly effective at modeling complex nonlinear relationships thanks to kernel functions [32]. In this study, the RBF kernel is particularly well-suited to the nonlinear relationships between climate and malaria. These three approaches were selected because of their ability to handle complex interactions among explanatory variables and to model potentially nonlinear relationships in spatiotemporal malaria data.

Spatio - Temporal Data Partitioning and Validation

To rigorously build and evaluate the predictive models, the dataset covering the period 2016–2022 was structured into a strict chronological hierarchy tailored to epidemiological surveillance. This approach follows recommendations for forecasting studies, where model development and evaluation should respect the temporal ordering of observations to avoid information leakage and to mimic real-world prediction settings [21, 33].

I:Training set (2016–2020): Combined with regional block cross-validation to capture spatial-temporal dependencies and tune hyperparameters. II: Independent temporal test set (2021): Utilized for residual diagnostics and spatial autocorrelation analyses. III: Independent validation set (2022): Served as the final out-of-sample temporal horizon to replicate real-world forecasting conditions, is considered the gold standard for assessing model transportability because it evaluates predictive performance under conditions that closely resemble future operational deployment [34, 35].

Spatial dependence may inflate predictive performance when neighbouring observations are simultaneously represented in training and testing subsets. Spatial blocking strategies have therefore been recommended to obtain more realistic estimates of model performance in ecological and epidemiological prediction studies [36, 37].

Crucially, to eliminate data contamination and avoid optimistic bias, all hyperparameter tuning and feature selection steps were performed exclusively within the training dataset using time-series cross-validation folds. The test and validation sets were strictly held out until final performance evaluation. Restricting feature selection and hyperparameter tuning to the training set is considered essential to prevent data leakage, which may otherwise lead to overly optimistic estimates of predictive performance and poor generalizability to unseen data [38, 39].

Hyperparameter Tuning

Hyperparameters for RF, XGBoost, and SVM were optimized using a time-series cross-validation grid search evaluating a total of 90 distinct parameter combinations (36 for Random Forest, 27 for XGBoost, and

27 for SVM) [30, 40, 41]. The search spaces, evaluation grids, and optimal configurations are detailed in Table 1.

Table 1 Hyperparameter tuning space, evaluation grid, and optimal configurations for Random Forest, XGBoost, and Support Vector Machine models [42].

Models	Hyperparameters	Search Space / Tested Grid	Optimal Value
Random Forest	Number of trees	200, 500, 1.000	500
	Variables per split (*mtry*)	3,4,5,6,7,8,9,10	5
	Min. terminal node size (*min_n*)	2,5,10	5
XGBoost	Number of boosting iteratior	200, 500, 1.000	500
	Maximum tree depth	3,6,9	6
	Learning rate	0.01, 0.05, 0.10	0.05
SVM (RBF)	Cost penalization (*C*)	0.1, 1.0, 10.0	1.0
	Kernel width (*rbf_sigma*)	0.01, 0.05, 0.10, 0.50	0.10

Model Evaluation and Validation Metrics

Model performance was evaluated and validated using several metrics: **RMSE** (Root Mean Square Error): Measures the average difference between observed and predicted values, heavily penalizing large errors and outliers.

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

MAE (Mean Absolute Error) Quantifies the average absolute error without directional bias, offering less sensitivity to outliers than RMSE.

$$MAE = \left(\frac{1}{n}\right) * \Sigma\,|y_i - \hat{y}_i|$$

R^2 (coefficient of determination) Represents the proportion of total variance in malaria incidence explained by the model, ranging up to 1.

Residual Diagnostics and Spatial Independence

To test statistical assumptions and residual autocorrelation, comprehensive diagnostics were performed on model errors (observed minus predicted values). Residual stationarity was assessed using Augmented Dickey-Fuller and KPSS tests [43, 44]. Temporal autocorrelation was tested via the Ljung-Box test (12-month lag) and

Durbin-Watson statistics [21], while normality and homoscedasticity were verified using the Shapiro-Wilk and Breusch-Pagan tests, respectively. Furthermore, to evaluate spatial dependence across health regions, Moran's *I* test was computed on region-aggregated residuals using a nearest-neighbor spatial weights matrix supplemented by Monte Carlo simulations [45]. All analytical workflows were executed in R version 4.4.3 using the tidymodels framework, maintaining a fixed random seed (set.seed(123)) to ensure complete reproducibility of data partitions, cross-validation folds, and optimization steps [42].

Statistical Analysis and Forecast Comparison

To formally assess whether differences in predictive performance among the machine learning and baseline models were statistically significant, the **Diebold–Mariano test** was conducted using out-of-sample forecast errors from the independent validation dataset (2022). This test evaluates the null hypothesis of equal predictive accuracy between two competing models by comparing their loss differentials (using squared or absolute error losses), accounting for potential serial correlation in multi-step forecast errors. All pairwise Diebold–Mariano comparisons were executed in R, with statistical significance evaluated at a threshold of $p < 0.05$ [46].

Results

The comparative evaluation across the 2021 test dataset and the 2022 independent validation dataset revealed that machine learning frameworks effectively capture the seasonal transmission dynamics of malaria in Guinea-Bissau, characterized by incidence increases during the rainy season and prominent peaks between August and October. Performance varies by region: it is higher in high-incidence areas, where the models accurately capture fluctuations, and more limited in low-transmission areas, where greater variability is observed. A comparative evaluation of conventional time-series models (SARIMA, ETS, and GLMM with Splines) and machine learning algorithms (Random Forest, Support Vector Machine, and XGBoost) was conducted across the test dataset (2021) and the independent out-of-sample validation dataset (2022).

Table 2 Comparative Predictive Performance of Statistical and Machine Learning Models for Malaria Incidence Forecasting with 95% confidence intervals estimated using regional block bootstrapping (1,000 replications).

Type	Models	Data test			Validation set		
		RMSE	MAE	R^2	RMSE	MAE	R^2
Reference	SARIMA	2.56	1.97	0.703	2.51	1.97	0.74
		(1.51–3.38)	(1.11–2.92)	(0.48–0.97)	(1.51–3.38)	(1.11–2.92)	(0.48–0.97)
Reference	ETS	2.46	1.92	0.726	2.41	1.91	0.75
		(1.49–3.30)	(1.09–2.82)	(0.47–0.97)	(1.49–3.30)	(1.09–2.82)	(0.47–0.97)
Reference	GLMM + Splines	2.40	1.56	0.741	2.30	1.56	0.82
		(0.97–3.58)	(0.69–2.68)	(0.63–0.98)	(0.97–3.58)	(0.69–2.68)	(0.63–0.98)
ML	RF	2.49	1.92	0.77	1.39	1.13	0.90
		(1.42–3.44)	(1.09–2.87)	(0.51–0.98)	(0.92–1.80)	(0.65–1.62)	(0.79–0.98)
ML	SVM	2.65	2.02	0.71	1.96	1.77	0.90
		(1.35–3.73)	(1.02–3.01)	(0.40–0.97)	(1.38–2.47)	(1.29–2.28)	(0.80–0.96)
ML	XGBoost	2.27	1.74	0.78	1.57	1.22	0.91
		(1.28–3.15)	(0.89–2.63)	(0.52–0.97)	(0.87–2.24)	(0.65–1.85)	(0.82–0.98)

During the 2021 test period, XGBoost achieved the highest performance among evaluated approaches with an R^2 of 0.78 (95% CI: 0.52 – 0.97, RMSE = 2.27, MAE = 1.74), followed closely by Random Forest R^2 = 0.77, RMSE = 2.49, MAE = 1.92). Among traditional baselines, GLMM with Splines exhibited robust performance (R^2 = 0.74, RMSE = 2.40) (Table 2).

During independent temporal validation on the 2022 dataset, all machine learning models demonstrated improved performance. XGBoost achieved the best overall performance among machine learning approaches (R^2 = 0.91, 95% CI: 0.82 – 098; RMSE = 1.57; MAE=1.22). Random Forest and SVM also showed strong generalization capacity, reaching R^2 values of 0.90. Classical time-series models remained competitive (SARima R^2 = 0.74, R^2, ETS R^2 = 0.75 and GLMM R^2 = 0.82).

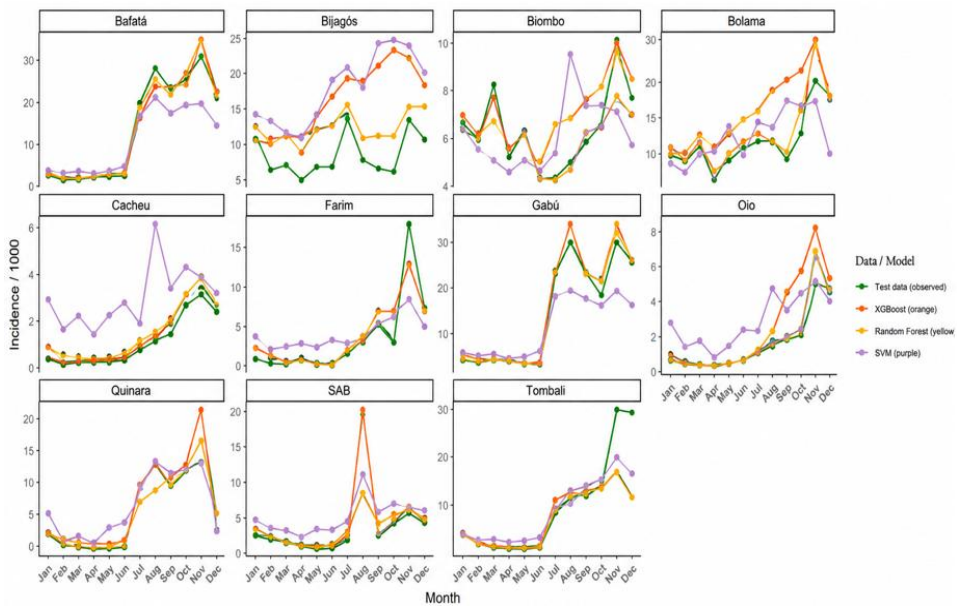


Fig. 4. Comparison of observed and predicted malaria incidence on the dataset (2021). Monthly malaria incidence by XGBoost (orange), Random Forest (yellow), and Support Vector Machine (SVM) (purple) compared with the observed incidence (test data, green) acros the health regions of Guinea–Bissau. Each subplot represent a region. The y-axix shows malaria incidence per 1,000 population, and the x-axis indicates the months of the year.

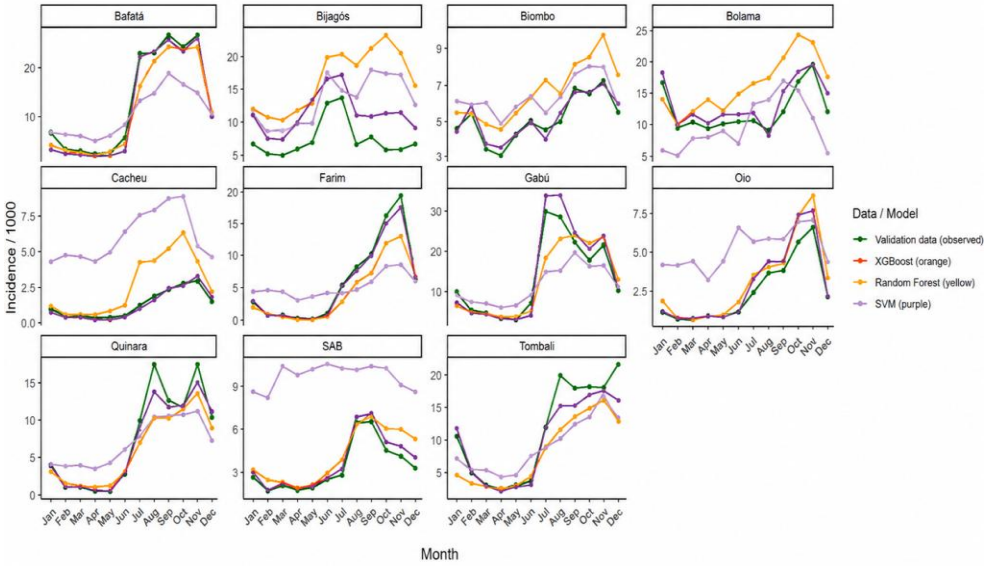


Fig. 5. Comparison of observed and predicted malaria incidence on the validation dataset (2022). Monthly malaria incidence by XGBoost (orange), Random Forest (yellow), and Support Vector Machine (SVM) (purple) compared with the observed incidence (validation data, green) acros the health regions of Guinea–Bissau. Each subplot represent a region. The y-axix shows malaria incidence per 1,000 population, and the x-axis indicates the months of the year.

Figures 4 and 5 illustrate the monthly observed and predicted malaria incidence for 2021 and 2022, respectively. All machine learning models successfully captured key seasonal patterns and rainy-season peaks (July–November). XGBoost outperformed other models, closely

matching peak timing and magnitude, whereas Random Forest showed intermediate accuracy and SVM tended to underestimate high-incidence periods. These consistent findings across both validation periods demonstrate the robustness and generalizability of the machine learning early warning approach.

Temporal Structure and Autocorrelation Dynamics

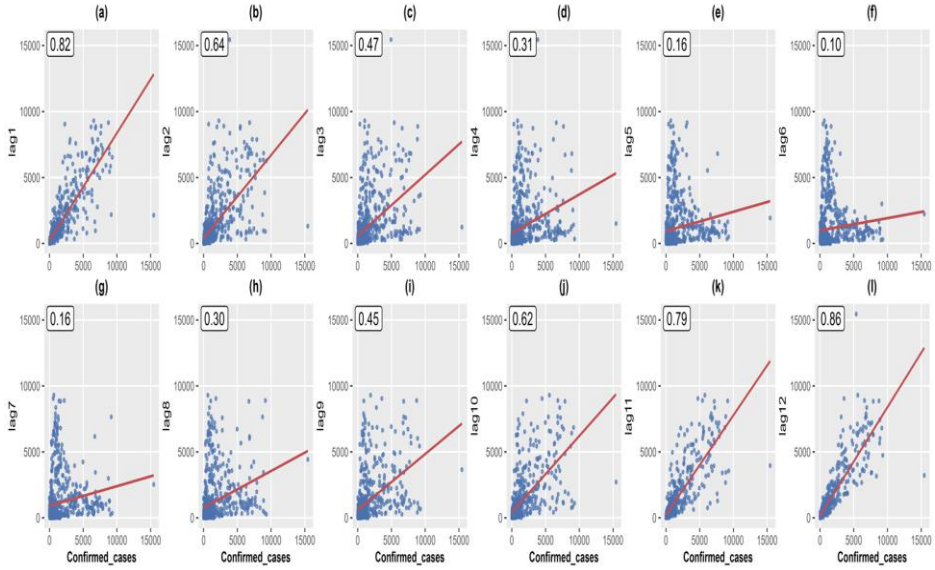


Fig. 6. Autocorrelation analysis of confirmed malaria cases across temporal lags. Subfigures (a) to (l) illustrate the bivariate relationship between monthly confirmed malaria cases and their respective time lags, from the 1st month time lag (a) to the 12th month time lag (l). Red trend lines represent non-linear fitted LOESS curves. Embedded box values in each panel denote the corresponding Pearson correlation coefficient (r).

Our temporal analysis revealed that historical case incidence (autoregressive lags) served as the most dominant and influential predictor across the machine learning models. Short-term temporal lags (particularly lag 1) captured immediate transmission momentum and persistent reservoir dynamics, while the 12-month lag effectively anchored the annual cyclical recurrence of malaria (Fig. 6). This high predictive weight of past incidence highlights the strong internal autoregressive inertia of malaria transmission systems, where preceding infection levels intrinsically dictate subsequent epidemiological trajectories.

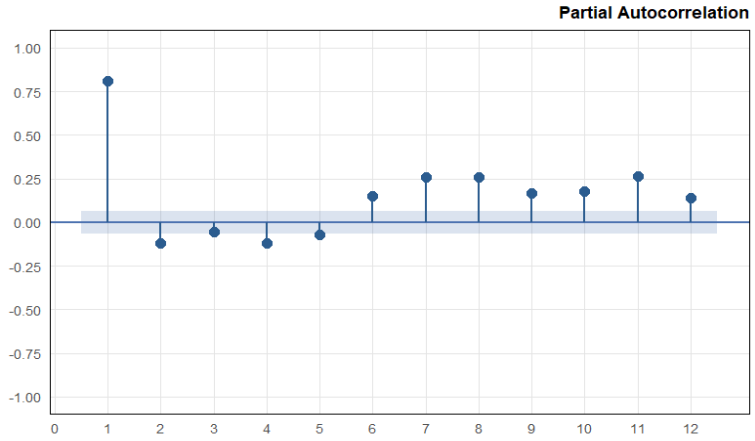


Fig. 7. Partial autocorrelation function (PACF) of monthly confirmed malaria cases across 12 temporal lags. The vertical spikes represent the partial correlation coefficients at each specific lag, with the shaded horizontal band indicating the 95% confidence interval for white noise. Significant spikes at early lags inform the direct autoregressive dependencies.

The partial autocorrelation function (PACF) of monthly confirmed malaria cases across 12 temporal lags is illustrated in Fig. 7. The vertical spikes represent the partial correlation coefficients at each specific lag, with the shaded horizontal band indicating the 95% confidence interval for white noise. A prominent and highly significant positive spike is observed at lag 1 (exceeding 0.75), indicating a strong direct autoregressive dependency from the immediately preceding month. Minor fluctuations are visible across subsequent intermediate lags, while spikes at longer lags remain largely within or close to the significance bounds. To assess whether the model residuals contain significant remaining autocorrelation, the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) of the residuals were examined (Fig.8). All estimated autocorrelation and partial autocorrelation coefficients remain well within the 95% confidence bounds (indicated by the dashed blue lines) across all evaluated lags. This indicates that the residuals are free from significant serial correlation, confirming that the model has successfully captured the linear temporal dependencies present in the data.

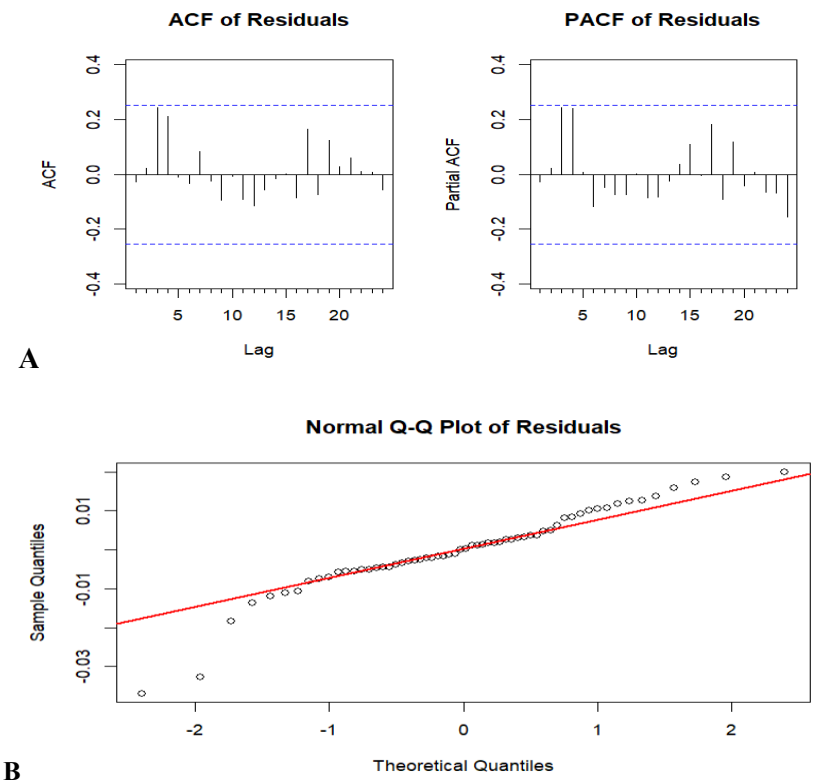


Fig. 8. (ACF/PACF & Q-Q-plots): Visual inspection of residual ACF/PACF plots showed that most autocorrelation coefficients remained within 95% confidence bounds, confirmed by Ljung-Box tests ($p > 0.05$). Q-Q plots showed residuals closely following theoretical normal distribution in central quantiles with minor tail deviations, consistent with Shapiro-Wilk test outcomes.

The absence of significant spikes beyond the confidence limits in both the ACF and PACF residual plots indicates that the model residuals approximate white noise. In time-series and epidemiological forecasting, verifying residual independence is a crucial step to validate that predictive performance is not biased by unmodeled temporal structures. As highlighted in standard time-series analytical frameworks (58, 59), uncorrelated residuals confirm that the temporal dynamics have been adequately accounted for by the modeling architecture.

To validate the distributional assumptions of the regression models, a normal Q-Q plot of the residuals was examined. The residual quantiles closely follow the theoretical normal distribution line across the central domain (between 1 and 1 on the theoretical quantiles), indicating a robust fit for the majority of the data points, with minor departures observed primarily at the lower tail. The evaluation of residual normality via the Q-Q plot indicates that while the model residuals approximate a normal distribution along the central quantiles, slight

deviations at the extremes suggest mild heavy-tailed behavior or minor skewness. In environmental and disease surveillance modeling particularly when analyzing aggregated monthly epidemiological counts such departures from strict normality are common due to overdispersion and zero-inflation inherent in public health data, as discussed in foundational works regarding generalized linear and non-linear modeling frameworks [28].

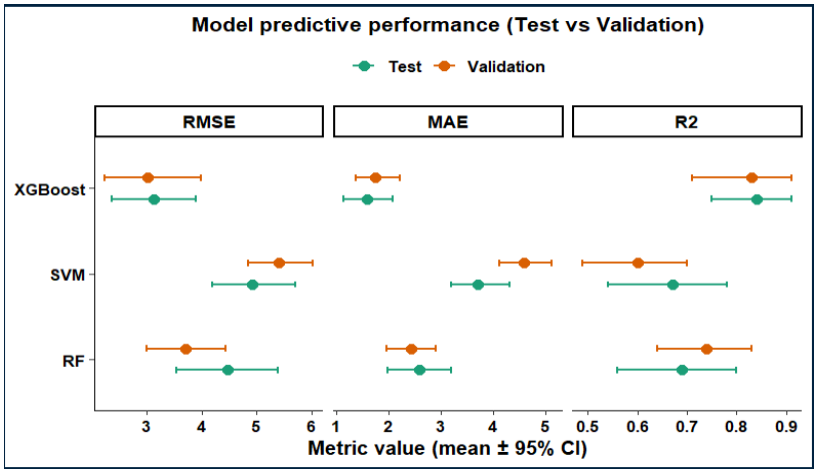


Fig. 9. Comparison of predictive performance of machine learning models on the test (2021) and validation (2022) datasets. The performance of XGBoost, RF, and SVM models was assessed using RMSE, MAE, and R^2. Points represent mean metric values, and horizontal bars indicate 95% confidence intervals.

The comparative performance metrics clearly demonstrate the superiority of the Extreme Gradient Boosting (XGBoost) architecture over Random Forest and Support Vector Machines in forecasting monthly malaria incidence (Fig. 9). The low error margins (RMSE and MAE) and high coefficient of determination $R^2 > 0.80$ achieved by XGBoost underscore its robust capacity to handle complex non-linear relationships and tabular environmental interactions. These findings align with recent machine learning applications in epidemiological surveillance [31], which emphasize that gradient boosting frameworks frequently surpass traditional benchmark models in capturing multi-variable disease dynamics and minimizing generalization error across unseen temporal folds.

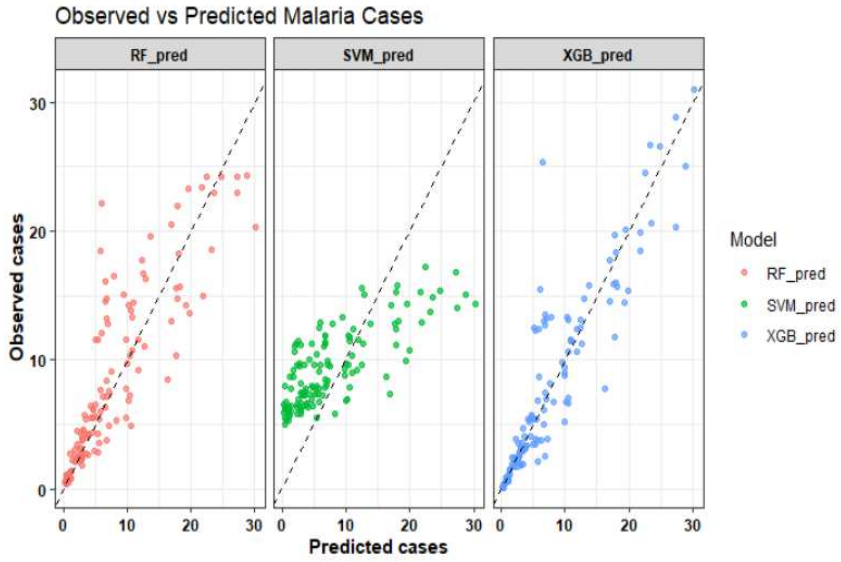


Fig. 10. Scatter plots displaying the relationship between observed malaria cases (y-axis) and predicted cases (x-axis) by the Random Forest (RF_pred), Support Vector Machine (SVM_pred), and Extreme Gradient Boosting (XGB_pred) models. The dashed diagonal line represents the perfect 1:1 agreement. Points closer to the line indicate higher prediction accuracy. Data points represent weekly aggregated incidence from 2016–2022.

A comparative evaluation of regression algorithms (Random Forest, SVM, and XGBoost) shows that XGBoost is clearly superior (Fig. 10). Analysis of the scatter plots (observed vs. predicted) indicates that XGBoost best minimizes the mean squared error and accurately captures the incidence peaks ($y \approx x$), whereas SVM exhibits a systematic bias toward underestimating extreme values [30, 31].

Forecast Accuracy Comparison Using Diebold–Mariano Tests

Diebold–Mariano tests were conducted on the independent validation dataset (2022) to assess whether differences in forecast accuracy between models were statistically significant. Although XGBoost achieved the best overall predictive performance (RMSE = 1.57, MAE = 1.22, $R^2 = 0.91$), no significant differences in forecast accuracy were detected relative to the competing models. For example, the comparison between XGBoost and SARIMA showed no significant difference (DM = 0.012, p = 0.991), and similar results were obtained for ETS, GLMM + Splines, Random Forest, and SVM. These findings suggest that, despite its superior point-estimate performance, XGBoost did not demonstrate statistically superior forecasting accuracy on the validation dataset. The limited validation period (12 monthly observations) may have reduced the statistical power of the tests.

Residual Diagnostics and Spatial Autocorrelation

The malaria incidence series satisfied stationarity assumptions, as supported by significant Augmented Dickey–Fuller tests (p = 0.010) and non-significant KPSS tests (p = 0.100). Residual autocorrelation diagnostics indicated that the temporal dependence structure was adequately captured, with non-significant Ljung–Box statistics for SARIMA ($p = 0.912$), ETS ($p = 0.612$), and GLMM + Splines ($p = 0.091$) (Table 3).

Table 3 Residual Diagnostic Tests for SARIMA, ETS, and GLMM Models.

Test	SARIMA	ETS	GLMM + Splines
Dickey-Fuller ADF p-value	0.0100	0.0100	0.0100
KPSS p-value	0.1000	0.1000	0.1000
Ljung-Box p-value	0.9124	0.6118	0.0905
Shapiro-Wilk p-value	0.0000	0.0081	0.3201
Durbin-Watson	0.0269	0.0592	1.2065
Breusch-Pagan p-value	NA	NA	0.0085

Spatial Autocorrelation (Moran's I)

To assess whether the spatial distribution of the model residuals exhibited remaining spatial dependence, Moran's I test under randomisation was performed using spatial weights (lw). The test indicated no statistically significant spatial autocorrelation in the mean residuals (Moran's I = 0.065, expected = -0.091, standard deviate = 0.660, p = 0.255). This finding was further corroborated by a Monte-Carlo simulation (1,000 permutations), which yielded a consistent outcome (statistic = 0.065, observed rank = 755, p = 0.245). Together, these results confirm that the model successfully captured the primary spatial structures, and the residuals are spatially independent, validating the assumption of spatial independence required for subsequent analyses.

Sensitivity Analysis (COVID-19 Disruption Period)

To evaluate whether models trained on pre-pandemic surveillance data maintained their predictive capacity during the pandemic disruption era, a sensitivity analysis was conducted using a training period of 2016–2019 and an evaluation period spanning 2020–2022.

Table 4 Sensitivity Analysis during the COVID-19 Period and Comparison with Principal Model.

Models	R^2 Main validation	R^2 Pré-COVID training validation	Relative change
RF	0.90	0.40	↓ 55.6%
SVM	0.90	0.46	↓ 48.9%
XGBoost	0.91	0.41	↓ 54.9%

The results revealed a substantial reduction in predictive accuracy when models calibrated on pre-pandemic baselines were applied to the COVID-19 period, with R^2 values dropping sharply from approximately 0.90 down to 0.40 – 0.46 across all machine learning architectures (RF dropping by 55.6%, SVM by 48.9%, and XGBoost by 54.9%) (Table 4). This notable performance decline underscores the vulnerability of forecasting models to systemic shocks and temporal distributional shifts.

Model Interpretation and Environmental Drivers (SHAP & PDPs)

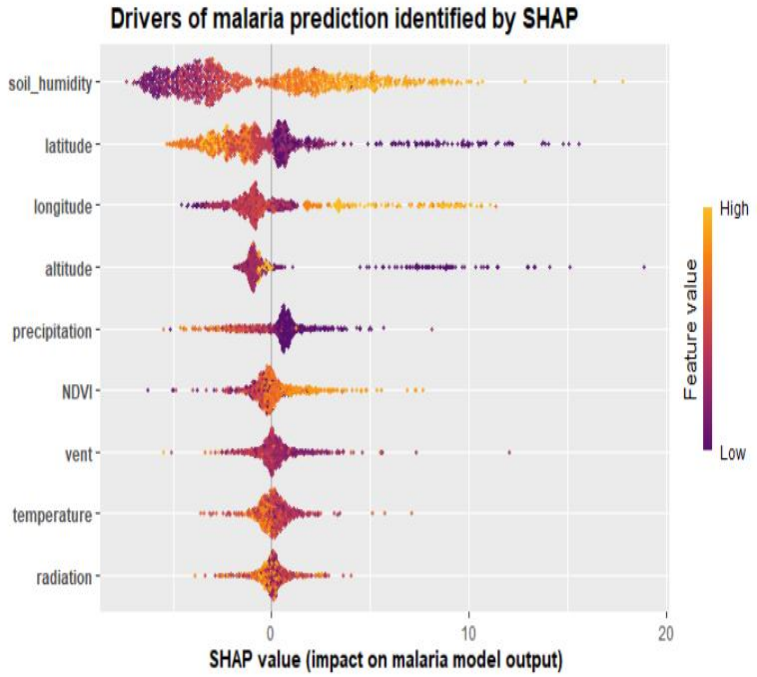


Fig. 11. SHAP summary plot of variable contributions to malaria incidence predictions from the XGBoost model. SHAP values on the x-axis represent the impact of each predictor on model output; positive values increase predicted malaria incidence, while negative values decrease it. Each point corresponds to an observation. The colour scale represents the original feature value, with low values shown in purple/blue and high values shown in orange/yellow, indicating whether higher or lower values of each predictor contribute to increased or reduced malaria predictions.

To interpret the black-box mechanics of the optimal XGBoost model, SHapley Additive exPlanations (SHAP) were computed (Fig. 11). The summary plot illustrates the distribution of feature impacts on model output, where SHAP values on the x-axis denote the magnitude and direction of each predictor's contribution to model predictions. Soil moisture (*soil_humidity*) emerged as the most influential environmental predictor, with high values (yellow/orange dots) predominantly yielding positive SHAP values thus driving higher predicted malaria incidence [29, 47]. Geographic coordinates (latitude and longitude) and altitude followed closely in importance, reflecting substantial spatial heterogeneity captured by the model. Vegetation indices (NDVI) and precipitation also showed positive predictive contributions at high values, whereas meteorological factors such as wind speed, temperature, and solar radiation exhibited distributions more concentrated around zero, indicating comparatively weaker contributions to model forecasts.

Importantly, SHAP values quantify predictive importance within the joint distribution of the training data and should not be interpreted as evidence of causal effects. As illustrated in the conceptual directed acyclic graph (DAG) presented in Supplementary Figure S1, climatic, environmental, spatial, socioeconomic, and health-system factors are interconnected through complex pathways that may involve unmeasured confounding and indirect relationships. Therefore, variables identified as highly influential by SHAP should be regarded as predictors associated with malaria incidence rather than direct determinants of malaria transmission [48, 49].

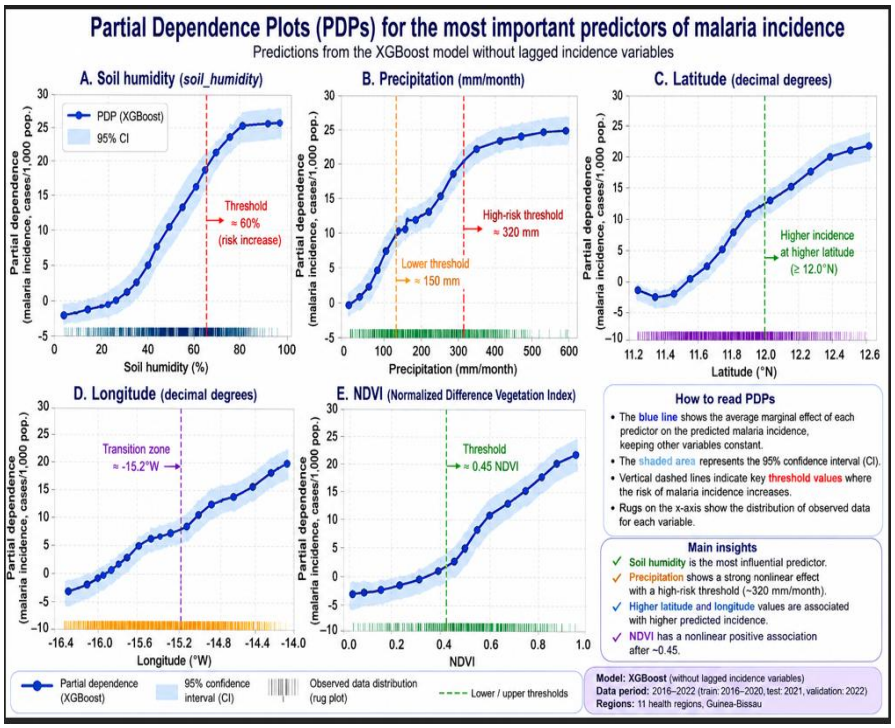


Fig. 12. Partial dependence plots showing the marginal effects of the main environmental and spatial predictors on malaria incidence predicted by the XGBoost model. The curves represent the average marginal effect of each predictor on malaria incidence while holding all other variables constant. Shaded areas represent 95% confidence intervals estimated using regional block bootstrap (1,000 replications). Vertical dashed lines indicate approximate risk thresholds identified from the PDP analysis

Partial dependence plots (PDPs) were used to further interpret the marginal effects of the main environmental and spatial predictors identified by the SHAP analysis (Fig. 12). The curves illustrate non-linear relationships between several environmental factors and the predicted incidence of malaria, holding all other variables constant. Soil moisture (*soil_humidity*) emerged as the most influential environmental determinant, displaying a sharp increase in risk once values reach approximately 60%, reflecting conditions highly favorable to the formation and persistence of larval habitats. Precipitation exhibited a strong non-linear effect, with a lower risk threshold at approximately 150 mm/month and a high-risk threshold near 320 mm/month, suggesting that heavy rainfall indirectly promotes transmission by increasing aquatic breeding sites. Similarly, the Normalized Difference Vegetation Index (NDVI) showed a positive non-linear association with a critical threshold at approximately 0.45, indicating that denser vegetation cover provides conducive ecological conditions for mosquito development. Furthermore, geographic coordinates specifically latitude (with higher incidence associated above ≥ 12.0ºN)

and longitude (identifying a transition zone around – 15.2ºW) highlight significant spatial heterogeneity in malaria risk across the health regions of Guinea-Bissau. These relationships confirm that environmental and spatial factors continue to make a substantial contribution to predictive performance even when excluding lagged incidence variables.

Model Forecast Based on a Heat Map

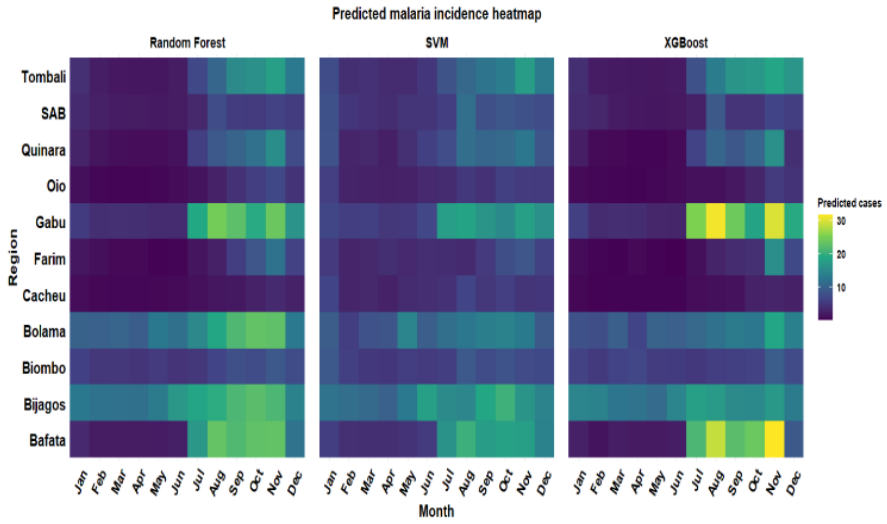


Fig. 13. Heat map of malaria incidence predicted by the XGBoost model by month and region in Guinea-Bissau. The regions are plotted on the y-axis, while the months are plotted on the x-axis. Each cell corresponds to the predicted malaria incidence for a given region in a given month.

The heat map of predicted malaria incidence by month and health region (Fig. 13) reveals pronounced spatial heterogeneity and marked temporal seasonality across Guinea-Bissau. A prolonged period of low incidence is observed across all models from January through June, followed by a sharp, uniform increase coinciding with the rainy season from July through November. Regarding regional heterogeneity, the models consistently identify Gabu and Bafata as areas of high transmission intensity (with predicted incidence exceeding 30 cases), whereas regions such as Cacheu and Biombo exhibit a lower disease burden. This distinct spatial distribution underscores how local environmental conditions and regional characteristics strongly influence epidemic susceptibility.

Discussion

The findings demonstrate the superior predictive performance of tree-based machine learning architectures particularly XGBoost [31], over conventional linear models and Support Vector Machines (SVM) in forecasting malaria incidence across Guinea-Bissau. This superiority stems from XGBoost's capacity to model complex, non-linear interactions between meteorological covariates and parasite incidence [50]. In contrast, kernel approaches like SVM tend to smooth out extreme epidemiological peaks, while classical stationary models struggle with climatic volatility and non-linear lags, as heavily documented in time-series forecasting frameworks [21]. These results align with recent machine learning applications in West African vector surveillance, highlighting that non-linear algorithms effectively handle multi-variable disease dynamics [10].

Historical incidence emerged as a core predictor, reflecting the continuity of human infection reservoirs and local vector population persistence from one month to the next. The sharp truncation observed in the PACF analysis (Fig. 8) at lag 1 confirms that short-term autoregressive memory captures the momentum of transmission. As noted in comparable regional studies and supported by generalized linear modeling principles for count data [51], autoregressive components account for the largest share of short-term predictive accuracy because past case counts encapsulate unmeasured localized behavioral, operational, and micro-environmental factors that static baseline covariates fail to capture.

Beyond autoregressive terms, meteorological and environmental drivers played a mechanistic role by establishing boundary conditions that allowed models to anticipate shifts in transmission intensity months in advance, reinforcing established frameworks on climate-driven early warning systems [17, 52]. Soil moisture outranked direct precipitation in SHAP feature attribution, indicating that hydrological retention serves as a more reliable proxy for persistent vector breeding habitats and larval development micro-climates than surface rainfall alone an approach robustly supported by ecological and spatial statistical modeling extensions [53, 54]. Furthermore, the high importance of spatial coordinates (latitude, longitude, and altitude) underscores

significant spatial structuring and regional heterogeneity across Guinea-Bissau's health regions. As highlighted in spatial epidemiology literature, spatial coordinates frequently absorb unmeasured environmental and socioeconomic gradients rather than acting as direct biological mechanisms. Non-linear marginal effects from Partial Dependence Plots further substantiate these dynamics, identifying critical operational thresholds such as soil moisture surpassing 60% and precipitation exceeding ~320 mm which trigger sudden expansions of *Anopheles* larval habitats [17, 55].

Spatio-temporal validation maps effectively captured the stark epidemiological contrast between the low-incidence dry season (January to June) and the high-incidence monsoonal rainy season (July to November), identifying inland zones, Gabú and Bafatá as persistent high-transmission hotspots compared to northern or coastal areas. However, model generalizability declined during the 2020–2022 evaluation window. This degradation highlights how external disruptions such as alterations in healthcare-seeking behavior, reporting anomalies, and supply chain interruptions during public health emergencies can induce domain shift in data-driven surveillance systems [56]. As emphasized in broader literature on responsible machine learning deployment in healthcare [26], operational early-warning systems must incorporate real-time resilience metrics and adaptive recalibration protocols to maintain robustness against structural anomalies.

Crucially, these findings carry direct operational value for the National Malaria Control Program of Guinea-Bissau. By accurately forecasting transmission peaks months in advance, the validated XGBoost framework shifts public health planning from reactive responses to proactive, targeted interventions. Health authorities can leverage these granular outputs to optimize the pre-emptive deployment of vector control measures (such as seasonal chemoprevention and targeted indoor residual spraying) and manage antimalarial supply chains prior to peak transmission. Finally, the observed lower predictive accuracy in low-transmission settings underscores the necessity of adapting surveillance strategies to local epidemiological strata, reinforcing that granular, data-driven frameworks are vital to support pre-elimination efforts across diverse regional contexts.

Limitations

Several limitations of this study should be acknowledged. First, a substantial decline in predictive performance during the sensitivity analysis indicated that COVID-19 related disruptions significantly affected the stability and reliability of surveillance data during 2020–2022. Second, certain unmeasured variables such as entomological indicators, vector control coverage (insecticide-treated nets and indoor residual spraying), population mobility, land-use characteristics, and healthcare access indicators were not consistently available throughout the study period. Third, the analyses relied on data aggregated at regional and monthly levels, which may obscure local micro-environmental conditions and finer spatial-temporal heterogeneity. Finally, SHAP values reflect predictive associations and feature contributions within the observed dataset rather than definitive causal biological mechanisms, meaning that geographic variables like latitude and longitude largely absorb unmeasured ecological or socioeconomic gradients.

Conclusion

Our study on predictive modeling of malaria incidence in Guinea-Bissau highlights the critical value of advanced machine learning approaches in public health surveillance. By systematically comparing Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost), we demonstrated that XGBoost achieves superior performance in forecasting monthly incidence. The results indicate that these models effectively capture spatio-temporal trends, enabling health authorities to anticipate peak transmission periods between August and November and improve intervention responsiveness across persistent risk areas. Furthermore, SHAP interpretability revealed that soil moisture acts as a primary predictor, complemented by essential meteorological, geographic, and environmental covariates. Overall, these findings underscore the potential of data-driven predictive frameworks to support decision-making, resource allocation, and targeted control efforts in malaria-endemic settings.

Future work should incorporate operational intervention variables such as the distribution of insecticide-treated bed nets, seasonal malaria chemoprevention, and population mobility metrics to further refine

predictive accuracy. Additionally, exploring advanced deep learning architectures, including Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) models, or attention-based mechanisms, could better capture complex temporal dependencies in malaria transmission dynamics.

Abbreviations

IC	Confidence interval
CV	Cross-validation
EWS	Early warning system
GADM	Global Administrative Areas
GB	Guinea-Bissau
GLDAS	Global Land Data Assimilation System
LSTM	Long Short-Term Memory
MAE	Mean absolute error
ML	Machine Learning
NASA	National Aeronautics and Space Administration
NDVI	Normalized Difference Vegetation Index
NMCP	National Malaria Control Program
PDP	Partial Dependency graph
R^2	Coefficient of determination
RBF	Radial Basis Function
RDT	Rapid Diagnostic Tests
RF	Random Forest
RMSE	Root Mean Squared Error
RNNs	Recurrent Neural Networks
SAB	Sector Autonomous of Bissau
SHAP	Shapley's Additive Explanations
SRTM	Shuttle Radar Topography Mission
SVM	Support Vector Machine
VIF	Variance Inflation Factor
WHO	World Health Organization
XGBoost	Extreme Gradient Boosting

Declarations

Ethics approval and consent to participate

This study used secondary, anonymised, aggregated data at the sanitary-area level from the national DHIS2 platform, obtained through the WAMCAD project with authorisation from the Ministry of Public Health of Guinea-Bissau (PNLP, ref. 2/2025, 20/01/2025). Data access was granted via institutional data-sharing arrangements within the WAMCAD collaborative project. As no primary data collection or identifiable information was involved, formal ethics committee review was not required, consistent with CIOMS Guideline 12 and the Declaration of Helsinki.

Consent for publication

Not applicable. This study does not contain any individual data, images or information that could identify the participants.

Availability of data and materials

The malaria clinical data used in this study originate from the National Malaria Control Programme (NMCP) of Guinea-Bissau. Demographic data are maintained by the National Institute of Statistics (INE). Weather and environmental datasets are publicly available from the NASA POWER platform (https://power.larc.nasa.gov/). Geographic coordinates and shapefiles can be downloaded from GADM (https://gadm.org/).
Model scripts and code can be accessed in the public repository at https://github.com/Tchalino/malaria-ml-guinea-bissau.

Competing interests

The authors declare that they have no conflict of interest.

Funding

This study was funded by a grant from the Gates Foundation, GRANT AGREEMENT Investment ID (INV – 047051) The funders had no role or influence on the design and interpretation of the data collected, as well as in writing the manuscript.

Authors contributions

- Study design and methodology: Marcelino Correia NANQUE, Mouhamad M. ALLAYA, Mor Absa LOUM and Mamadou BOUSSO
- Data collection: Marcelino Correia NANQUE
- Data analysis and interpretation: Marcelino Correia NANQUE, Mouhamad M. ALLAYA, Mamadou BOUSSO
- Manuscript writing – original draft: Marcelino Correia NANQUE
- Critical revision of the intellectual content: Mouhamad M. ALLAYA, Mor Absa LOUM
- Supervision: Jean Louis Abdourahim NDIAYE
- Final approval of the manuscript: All authors have read and approved the final version of the manuscript. They agree to take responsibility for its content.

Acknowledgements

We would like to thank the Bill and Melinda Gates Foundation, West Africa Mathematical Capacity Development and Guinea-Bissau's National Malaria Control Program for providing access to malaria surveillance data. The authors also thank the supervisors and thesis advisors for their valuable scientific guidance and constructive feedback throughout this study.

Author details

[1]École Doctorale Développement Durable Et Société - ED2DS, Pôle Sciences de la Vie et de la Santé (SVS), Université Iba Der Thiam de Thiès, Thiès, Sénégal

[2]UFR des Sciences Économiques et Sociales, Université Iba Der Thiam de Thiès, Thiès, Sénégal

[3]UFR des Sciences de l'Ingénieur, Université Iba Der Thiam de Thiès, Thiès, Sénégal

[4]Departemento de Seguimento e avaliação, Programa Nacional de Luta contra Paludismo (PNLT), Bissau, Guiné-Bissau

[5]Laboratoire d'Ecologie Vectorielle et Parasitaire, Faculté des Sciences et Techniques, Université Cheikh Anta Diop De Dakar, Dakar, Sénégal

[6]UFR des Sciences Appliquée et Technologies, Université Gaston Berger, Saint-Louis, Sénégal

[7]Instituto Nacional de Saúde Pública, Bissau, Guiné-Bissau

[8]École Doctorale Développement Durable Et Société - ED2DS, Pôle Sciences, Technologies, Mathématiques et Ingénierie (STMI), Université Iba Der Thiam de Thiès, Thiès, Sénégal

[9] Projeto Saúde Bamdim, Bissau, Guiné-Bissau

[10]Service Parasitologie Mycologie, UFR Santé Thiès, Université Iba Der Thiam de Thiès, Thiès, Sénégal

References

1. WHO World Health Organization (WHO). World malaria report; 2023.https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2023.

2. Satoguina J, Walther B, Drakeley C, Nwakanma D, Oriero EC, Correa S, et al. Comparison of surveillance methods applied to a situation of low malaria prevalence at rural sites in The Gambia and Guinea Bissau. Malar J. 2009;8:274. https://doi.org/10.1186/1475-2875-8-274.

3. Ryan SJ, Carlson CJ, Mordecai EA, Johnson LR. Global expansion and redistribution of Aedes-borne virus transmission risk with climate change.PLoSNeglTropDis.2019;13:e0007213. https://doi.org/10.1371/journal.pntd.0007213.

4. National Malaria Control Program. World Health Organization. World Malaria Report 2023. Geneva: World Health Organization; 2023. Available from: World Malaria Report 2023.

5. Ministry of Public Health of Guinea-Bissau. National Strategic Plan: Roll Back Malaria in Guinea-Bissau 2006–2010. National Malaria Control Programme; 2006.

6. Ministry of Public Health of Guinea-Bissau. Republic of Guinea-Bissau. First Biennial Update Report (BUR1) of Guinea-Bissau under the United Nations Framework Convention on Climate Change (UNFCCC). Bissau: Ministry of Environment and Sustainable Development.Availablefrom: https://unfccc.int/sites/default/files/resource/FINAL_GNB_BUR1.pdf. National Malaria Control Programme; 2021.

7. Rodrigues A, Schellenberg JA, Kofoed P-E, Aaby P, Greenwood B. Changing pattern of malaria in Bissau, Guinea Bissau. Trop Med Int Health TM IH. 2008;13:410–7. https://doi.org/10.1111/j.1365-3156.2008.02016.x.

8. Silva R, Lopes LF, Rodrigues A, Arez AP, Medeiros MM. Assessing the burden of submicroscopic Plasmodium infections in a pre-elimination malaria setting in sub-Saharan Africa, Guinea-Bissau. Malar J. 2024;23:316. https://doi.org/10.1186/s12936-024-05138-z.

9. Ursing J, Rombo L, Rodrigues A, Aaby P, Kofoed P-E. Malaria transmission in Bissau, Guinea-Bissau between 1995 and 2012: malaria resurgence did not negatively affect mortality. PloS One. 2014;9:e101167. https://doi.org/10.1371/journal.pone.0101167.

10. Bakhoum LN, Loum MA, Allaya MM, Gbaguidi GJ, Ndiaye K, Ly AY, et al. Causal Structural Equation Modeling of Climatic and Environmental Factors Influencing Malaria Incidence in Southeastern Senegal.IntJStatMedRes.2026;15:186–200. https://doi.org/10.6000/1929-6029.2026.15.17.

11. Venkatesan P. WHO world malaria report 2024. Lancet Microbe. 2025;:101073. https://doi.org/10.1016/j.lanmic.2025.101073.

12. World Health Organization. Malaria 2024: Guinea-Bissau Country Profile. Geneva: World Health Organization; 2024. Available from: WHO Guinea-Bissau Malaria Country Profile 2024. https://www.who.int/publications/m/item/malaria-2024-gnb-country-profile.

13. Caminade C, Kovats S, Rocklov J, Tompkins AM, Morse AP, Colón-González FJ, et al. Impact of climate change on global malaria distribution. Proc Natl Acad Sci USA. 2014;111:3286–91. https://doi.org/10.1073/pnas.1302089111.

14. Bhatt S, Weiss DJ, Cameron E, Bisanzio D, Mappin B, Dalrymple U, et al. The effect of malaria control on Plasmodium falciparum in Africabetween2000and2015.Nature.2015;526:207–11. https://doi.org/10.1038/nature15535.

15. Socioeconomic disparities remain important across regions of Guinea-Bissau according to the Harmonized Household Living ConditionsSurvey2021–2022. https://microdata.worldbank.org/index.php/catalog/6274.

16. Nicholson SE. The West African Sahel: A Review of Recent Studies on the Rainfall Regime and Its Interannual Variability. Int Sch Res Not. 2013;2013:453521. https://doi.org/10.1155/2013/453521.

17. Thomson MC, Doblas-Reyes FJ, Mason SJ, Hagedorn R, Connor SJ, Phindela T, et al. Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. Nature. 2006;439:576–9. https://doi.org/10.1038/nature04503.

18. Mordecai E, Caldwell J, Grossman M, Lippi C, Johnson L, Neira M, et al. Thermal biology of mosquito-borne disease. Ecol Lett. 2019;22:1690–708. https://doi.org/10.1111/ele.13335.

19. Earth Science Data Systems N. NASA Earthdata. Earth Science Data Systems: Your gateway to NASA Earth observation data. 2026. Available from: https://www.earthdata.nasa.gov/. 2026.

20. Weese M, Martinez W, Megahed FM, Jones-Farmer LA. Statistical Learning Methods Applied to Process Monitoring: An Overview and Perspective.JQualTechnol.2016;48:4–24. https://doi.org/10.1080/00224065.2016.11918148.

21. Hyndman RJ, Athanasopoulos G. Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. 2nd ed. OTexts; 2018. Available from: https://otexts.com/fpp2/. OTexts; 2018.

22. Heuschen A-K, Lu G, Razum O, Abdul-Mumin A, Sankoh O, von Seidlein L, et al. Public health-relevant consequences of the COVID-19 pandemic on malaria in sub-Saharan Africa: a scoping review. Malar J. 2021;20:339. https://doi.org/10.1186/s12936-021-03872-2.

23. Despite continued impact of COVID-19, malaria cases and deaths remained stable in 2021. https://www.who.int/news/item/08-12-2022-despite-continued-impact-of-covid-19--malaria-cases-and-deaths-remained-stable-in-2021.

24. Molnar C. Interpretable machine learning: a guide for making black box models explainable. 2nd ed. 2022. Available from: https://christophm.github.io/interpretable-ml-book/.

25. Lundberg S, Lee S-I. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. arXiv. 2017.Available from: https://doi.org/10.48550/arXiv.1705.07874. 2017.

26. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for healthcare.NatMed. 2019;25:1337–40. https://doi.org/10.1038/s41591-019-0548-6.

27. Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography. 2013;36:27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x.

28. Zuur AF, Ieno EN, Elphick CS. A protocol for data exploration to avoid common statistical problems. Methods Ecol Evol. 2010;1:3–14. https://doi.org/10.1111/j.2041-210X.2009.00001.x.

29. Mordecai EA, Paaijmans KP, Johnson LR, Balzer C, Ben-Horin T, De Moor E, et al. Optimal temperature for malaria transmission is dramatically lower than previously predicted. Ecol Lett. 2013;16:22–30. https://doi.org/10.1111/ele.12015.

30. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32. doi: 10.1023/A:1010933404324. Random Forests. Mach Learn.

31. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.

32. Cortes C, Vapnik V. Support-Vector Networks. Mach Learn. 1995;20:273–97. https://doi.org/10.1023/A:1022627411411.

33. Kuhn M, Johnson K. Kuhn M, Johnson K. Applied predictive modeling. New York: Springer Science & Business Media; 2013. doi: 10.1007/978-1-4614-6849-3. Springer Science & Business Media; 2013.

34. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. Br J Cancer. 2015;112(1):e1-10.doi:10.1038/bjc.2014.639.
https://doi.org/10.1038/bjc.2014.639.

35. Steyerberg EW. Clinical prediction models: A practical approach to development, validation, and updating. 2nd ed. Cham: Springer; 2019. doi: 10.1007/978-3-030-16399-0.

36. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, et al. Cross-validation strategies for data with temporal, spatial, hierarchical,orphylogeneticstructure.
https://doi.org/10.1111/ecog.02881.

37. Valavi R, Elith J, Lahoz-Monfort JJ, Guillera-Arroita G. blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods Ecol Evol. 2019;10:225–32. https://doi.org/10.1111/2041-210X.13107.

38. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-basedscience.Patterns.2023;4.
https://doi.org/10.1016/j.patter.2023.100804.

39. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: Formulation, detection, and avoidance: ACM Transactions on Knowledge Discovery from Data: Vol 6, No 4. 2012. https://dl.acm.org/doi/10.1145/2382577.2382579.

40. Hastie, T., Tibshirani, R., & Friedman,. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition | SpringerNatureLink.2009.
https://link.springer.com/book/10.1007/978-0-387-84858-7.

41. Vapnik V. Vapnik VN. The Nature of Statistical Learning Theory. 2nd ed. New York: Springer; 2000. doi: 10.1007/978-1-4757-3264-1. Springer Science & Business Media; 2013.

42. Kuhn M, Silge J. Kuhn M, Silge J. Tidy Modeling with R: A Framework for Modeling in the Tidyverse. Sebastopol (CA): O'Reilly Media; 2022. Available from: Tidy Modeling with R (online book). O'Reilly Media, Inc.; 2022.

43. Dickey DA, Fuller WA. Distribution of the Estimators for Autoregressive Time Series with a Unit Root. J Am Stat Assoc. 1979;74:427–31. https://doi.org/10.1080/01621459.1979.10482531.

44. Kwiatkowski D, Phillips PCB, Schmidt P, Shin Y. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? J Econom. 1992;54:159–78. https://doi.org/10.1016/0304-4076(92)90104-Y.

45. Shapiro SS, Wilk MB. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika. 1965;52(3–4):591–611. doi:10.1093/biomet/52.3-4.591.

46. Diebold FX, Mariano RS. Diebold FX, Mariano RS. Comparing Predictive Accuracy. Journal of Business & Economic Statistics. 1995;13(3):253–263. doi:10.1080/07350015.1995.10524599. J Bus Econ Stat. 2002;20:134–44.

47. Parham PE, Michael E. Modeling the Effects of Weather and Climate Change on Malaria Transmission. Environ Health Perspect. 2010;118:620–6. https://doi.org/10.1289/ehp.0901256.

48. Textor J, Van Der Zander B, Gilthorpe MS, Liśkiewicz M, Ellison GTH. Robust causal inference using directed acyclic graphs: the R package'dagitty.'IntJEpidemiol.2017;:dyw341. https://doi.org/10.1093/ije/dyw341.

49. Greenland S, Pearl J, Robins JM. Causal Diagrams for Epidemiologic Research. Epidemiology. 1999;10:37–48.

50. Ly AY, Allaya MM, Loum MA, Gning LD, Sy O, Bakhoum LN, et al. Forecasting malaria incidence in the Southeast districts of Senegal using a machine learning approach. BMC Artif Intell. 2025;1:9. https://doi.org/10.1186/s44398-025-00007-4.

51. McCullagh, P., & Nelder, J. A. Generalized Linear Models by P. McCullagh, 2nd edition | 9780367488406, 9781351445849. 1989. https://www.perlego.com/book/1615873/generalized-linear-models-pdf.

52. Cox J, Abeku TA. Early warning systems for malaria in Africa: from blueprint to practice. Trends Parasitol. 2007;23:243–6. https://doi.org/10.1016/j.pt.2007.03.008.

53. Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM. Mixed Effects Models and Extensions in Ecology with R. New York: Springer; 2009. doi:10.1007/978-0-387-87458-6.

54. Molnar C, Casalicchio G, Bischl B. Interpretable Machine Learning -- A Brief History, State-of-the-Art and Challenges. 2020. p. 417–31. https://doi.org/10.1007/978-3-030-65965-3_28.

55. Mordecai EA, Caldwell JM, Grossman MK, Lippi CA, Johnson LR, Neira M, et al. Thermal biology of mosquito-borne disease. Ecol Lett. 2019;22:1690–708. https://doi.org/10.1111/ele.13335.

56. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, the Precise4Q consortium. Amann J, Blasimme A, Vayena E, Frey D, Madai VI; Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. 2020;20(1):310. doi:10.1186/s12911-020-01332-6. BMC Med Inform Decis Mak. 2020;20:310.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- NanqueetalSupplementaryMaterial.docx