

Original Article

Asian Pacific Journal of Tropical Medicine

apjtm.org

doi: 10.4103/apjtm.apjtm_681_25

Machine learning models to predict the risk of malaria outbreaks using balanced climatic and non-climatic data

Nwawudu Sixtus Ezenwa[1], Pedro Latorre-Carmona[2], J. Salvador Sánchez[1]✉

[1]*Institute of New Imaging Technologies, Department of Computer Languages and Systems, Universitat Jaume I, Campus Riu Sec s/n, 12071 Castellón de la Plana, Spain*

[2]*Computer Engineering Department, University of Burgos, Avda Cantabria s/n, 09006, Burgos, Spain*

ABSTRACT

Objective: To predict malaria outbreaks early and accurately using machine learning models when data presents challenges such as imbalance.

Methods: To predict the risk of malaria outbreaks, we applied six machine learning models—including naive Bayes, logistic regression, random forest, K-nearest neighbors, decision tree, and extreme gradient boosting—to a dataset of 8000 samples with 17 climatic and non-climatic features. To address class imbalance, we used the Synthetic Minority Over-sampling Technique algorithm. The models were evaluated using 10-fold cross-validation repeated 10 times to ensure robust validation.

Results: Extreme gradient boosting achieved the highest F1-score, improving from 0.926 on imbalanced data to 0.991 after balancing. While random forest, naive Bayes, and decision tree also improved their performance when balancing data with Synthetic Minority Over-sampling Technique (increasing F1-score from 0.893 to 0.931, from 0.765 to 0.962, and from 0.785 to 0.812, respectively), logistic regression and K-nearest neighbors did not show any improvement.

Conclusions: These findings demonstrate that integrating balanced climatic and non-climatic data with advanced machine learning can improve malaria outbreak risk prediction and support public health strategies.

KEYWORDS: Malaria outbreak prediction; Machine learning; Data imbalance; Public health; Synthetic Minority Over-sampling Technique

1. Introduction

Malaria is an infectious disease caused by a single-celled parasite of the genus *Plasmodium*, transmitted to humans by the bite of infected female *Anopheles* mosquitoes. After entering the human body, the parasites mature in the liver before infecting red blood cells,

Summary

Question: Which machine learning models can predict malaria outbreaks earlier and more accurately when the data is imbalanced?

Findings: Extreme gradient boosting and random forest achieved the best results in terms of F1-score due to their ability to integrate multiple weak classifiers, which reduces variance and improves generalization. Furthermore, the highest rates across all performance evaluation metrics were obtained with the balanced data, demonstrating the importance of resampling the dataset.

Meaning: The use of machine learning models along with algorithms to address data imbalances allows for accurate and early prediction of the risk of malaria outbreaks using both climatic and non-climatic data.

✉To whom correspondence may be addressed. E-mail: plcarmona@ubu.es

This is an open access article distributed under the terms of the Creative Commons Attribution-Non-Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

For reprints contact: reprints@medknow.com

©2026 *Asian Pacific Journal of Tropical Medicine* Produced by Wolters Kluwer-Medknow.

How to cite this article: Ezenwa NS, Latorre-Carmona P, Sánchez JS. Machine learning models to predict the risk of malaria outbreaks using balanced climatic and non-climatic data. Asian Pac J Trop Med 2026; 19(3): 129-137.

Article history: Received 27 November 2025 Revision 13 March 2026
Accepted 18 March 2026 Available online 24 March 2026

leading to the characteristic symptoms of the disease. Symptoms usually appear between 7 and 30 days after the mosquito bite, but sometimes they can take several months or even a year to develop. Malaria shares similar symptoms with other febrile illnesses such as influenza, pneumonia, dengue fever, or typhoid fever, including high fever, chills and shivering, profuse sweating, muscle and joint pain, fatigue, nausea, vomiting, and diarrhea. Because of this overlap in symptoms with those of many other diseases, a laboratory test (especially in the form of a microscopic examination or a rapid diagnostic test of a blood smear) is the only way to correctly diagnose malaria[1].

Malaria remains a major public health concern and is a leading cause of death in tropical and subtropical regions, particularly in sub-Saharan Africa[2]. According to the latest World Malaria Report from the World Health Organization (WHO), there were an estimated 263 million cases and 597000 malaria deaths worldwide in 2023, with an incidence of 60.4 cases per 1000 population at risk and a mortality rate of 13.7 per 100000[3]. This represents an increase of approximately 11 million cases compared to 2022, with Africa accounting for about 95% of the deaths. Although sub-Saharan Africa records the highest number of malaria cases, other high-risk regions include rural and forested areas of Southeast Asia (with the second highest burden of malaria cases globally), parts of Afghanistan, India and Pakistan, areas of the Amazon basin (Brazil, Peru, Colombia) and parts of Central America and the Caribbean (Guatemala, Honduras, Trinidad and Tobago and Martinique), as well as some Pacific islands (Papua New Guinea, Solomon Islands and Vanuatu)[4].

Despite advances in surveillance and control, early and accurate prediction of malaria outbreaks remains challenging, particularly due to issues such as data imbalance in predictive modeling. In this study, we address the impact of data imbalance on malaria outbreak prediction by applying and comparing six machine learning models—naive Bayes, logistic regression, random forest, K-nearest neighbors (KNN), decision tree, and extreme gradient boosting (XGBoost)—using a dataset of 8000 samples with 17 climatic and non-climatic features. We further evaluate the effectiveness of data balancing techniques in improving model performance, with the aim of enhancing early warning systems for malaria outbreaks.

Over the past decades, significant efforts have been made to predict malaria outbreaks incorporating geographic and climatic data. For example, Woube[5] and Teklehaimanot *et al.*[6] attempted to predict outbreaks using weather conditions, which are considered reliable indicators of malaria transmission. Rainfall levels, in particular, have been shown to influence not only the prevalence but also the transmission dynamics of malaria[7]. A 2005 study by Thomson *et al.*[8] demonstrated that rainfall and sea surface temperature are key factors for malaria early warning in Botswana. Similarly, Ndiaye *et al.*[9] found that malaria-related fatalities in the Niakhar region increased during the rainy season, regardless of temperature and humidity. Martineau *et al.*[10] showed that tropical climatic variability and associated sea surface temperature over the Pacific and Indian Oceans are valuable for predicting malaria in Limpopo, South Africa, up to three seasons ahead.

Christiansen-Jucht *et al.*[11] reported that temperature, precipitation, and humidity all affect the mosquito life cycle, including growth and development. According to Afrane *et al.*[12], vector larvae survive only when environmental conditions are favorable, specifically at moderate temperatures above 16 ℃, and perish when temperatures are lower or higher. When the climate supports mosquito survival, biting frequency increases.

Both machine learning algorithms and mathematical or statistical modeling approaches have been used for prediction. For example, Zinszer *et al.*[13] employed autoregressive integrated moving average models for weekly malaria forecasting, demonstrating the utility of using environmental (*e.g.*, rainfall, daytime and nighttime temperatures, and vegetation estimates) together with clinical variables (*e.g.*, type of antimalarial treatment). Khan *et al.*[14] compared the performance of decision trees, artificial neural networks, KNN, support vector machines with linear and radial kernels, logistic regression, XGBoost, and random forest integrating both climatic and non-climatic features to predict malaria outbreaks in each district of The Gambia.

Nkiruka *et al.*[15] developed a hybrid machine learning model for the early detection of malaria incidence, using climate variability across six countries of sub-Saharan Africa. To this end, they combined feature selection to identify relevant climatic factors, k-means clustering for outlier detection, and XGBoost for prediction, revealing that precipitation, temperature, and surface radiation contribute significantly to malaria outbreaks. Aheto *et al.*[16] employed three logistic regression models (lasso, ridge and elastic net) to identify the predictors of malaria prevalence in children under five years of age in Ghana. Harvey *et al.*[17] designed an early warning system for malaria epidemics based on a combination of a Gaussian process and a random forest regression model to predict the weekly number of malaria cases in a primary health center in Burkina Faso.

Okoronkwo *et al.*[18] analyzed five supervised machine learning methods (naive Bayes, support vector machines, linear regression, logistic regression, and KNN) using meteorological and malaria incidence data spanning from 2010 to 2020 to predict malaria outbreaks in Nigeria, showing an accuracy of 79.1% with naive Bayes and 75.5% with support vector machines. Ayoka and Nnadi[19] employed three regression models (random forest, multiple linear regression, and ridge regressor) to predict malaria prevalence using environmental, socioeconomic and behavioral covariates in Nigeria, demonstrating the usefulness of these models and consistent improvements as the amount of data increased.

Rahman and Shiddik[20] used spatial autocorrelation analysis to detect geographic clusters and hotspots of malaria, and combined XGBoost with explainable artificial intelligence (XAI) and causal artificial intelligence to predict malaria incidence and mortality, identify critical determinants of malaria, and establish causal relationships. Similarly, Monteiro *et al.*[21] evaluated the performance of six machine learning models to forecast weekly cases of malaria in several states of the Legal Amazon and further improved predictive accuracy by integrating k-means clustering for spatial analysis. Awe *et al.*[22] integrated ensemble-based machine learning models (random forest, AdaBoost, gradient boost, XGBoost, and CatBoost) and XAI techniques (LIME, SHAP, and permutation feature importance) to enhance the diagnosis accuracy of malaria on a dataset of 337 patients over a 4-week period and improve interpretability of the results.

Although there are multiple studies in which machine learning models are used to predict or model malaria outbreak risk and case incidence, only a few address the problem of data imbalance. A notable work is that of Lee *et al.*[23], which explored the effects of data imbalance on six machine learning models (support vector machines, random forest, multilayered perceptron, AdaBoost, gradient boosting, and CatBoost) using a dataset of patients with some parasitic disease, where the proportion of malaria patients was only 8%. They investigated whether applying Synthetic Minority Over-sampling Technique (SMOTE) improved the performance of the models, but found no significant improvements for this dataset. Mbaye *et al.*[24] compared six classical machine learning techniques (decision trees, random forest, naive Bayes, logistic regression, support vector machine, and artificial neural network) to predict malaria incidence using an imbalanced dataset with 21 083 patients in Senegal, of whom only 614 were diagnosed with malaria. To face the imbalance, they applied SMOTE and compared the machine learning algorithms with a baseline model based on a rapid diagnostic test.

Fleitas *et al.*[25] implemented a machine learning system to assess the risk of imported malaria in travelers. They experimented with a dataset comprising 4 443 anonymized traveler records, of which only 5% were diagnosed with malaria during or immediately after their trip. Therefore, they applied two methods—propensity score matching and SMOTE—to obtain training cohorts with different levels of data balance. The results showed that the best-performing models did not improve their predictive ability, regardless of the data balancing strategy used. To assess the risk of imported malaria in China, Yang *et al.*[26] employed a dataset of 27 088 records from the Chinese Center for Disease Control and Prevention, which exhibited a moderate data imbalance, so they addressed this critical challenge using SMOTE.

To predict malaria, Amosa *et al.*[27] performed a comparative analysis of four machine learning methods (logistic regression, random forest, KNN, and support vector machine). These methods were combined with three data balancing strategies (oversampling, under-sampling, and data augmentation) to improve model performance. The experimental results showed that using under-sampling with random forest achieved the highest accuracy. The main aim of this paper is to present a classification framework for early and accurately malaria prediction, in the presence of imbalance in the data distribution.

2. Methods

2.1. Dataset

The dataset used in this study was obtained from the Kaggle repository, which aggregates data from the South part of Nigeria (https://www.kaggle.com/datasets/scholarsway/malaria-outbreak-dataset). The dataset contains 8 000 samples, each represented by 19 environmental and epidemiological features and one target variable or class label, as shown in Table 1.

Table 1. Certain characteristics about the features of the dataset from the Kaggle repository.

Features	Type	Domain
Temperature level	Categorical	Low, high
Humidity	Categorical	Low, high
Precipitation	Categorical	No, yes
Travel patterns	Categorical	Low, high
Frequency outbreak	Categorical	Low, high
Mosquito species	Nominal	*Aedes aegypti*, *Anopheles gambiae*, *Culex quinquefasciatus*
Malaria vectors	Categorical	No, yes
Vectors infection rates	Categorical	Low, high
Healthcare availability	Categorical	No, yes
Healthcare accessibility	Categorical	No, yes
Malaria treatment	Categorical	Low, high
Land use	Categorical	Rural, urban
Mosquito breeding site	Categorical	No, yes
Stagnant water	Categorical	No, yes
Iot air quality	Categorical	Bad, good
Iot water quality	Categorical	Bad, good
Presence of vegetation	Categorical	No, yes
Treatment success rate	Categorical	Low, high
Outbreak risk	Target	Low, medium, high

Some machine learning models require numerical data to make predictions. However, all the attributes in the original dataset were categorical (for example, the possible values for the Land use attribute were “rural” and “urban”, or for the Humidity attribute were “low” and “high”), so it was first necessary to convert them into numerical data through one-hot encoding.

Initially, the target variable “Outbreak risk” had three possible values: low, medium, and high. However, upon closer examination of the data, it was found that the “high-risk” category was not represented as a distinct class but was instead combined with the “medium” category. As a result, the model could not reliably distinguish between “high-risk”

and "medium-risk" cases. Therefore, the two categories were merged, converting the problem from a three-class to a binary prediction task (low-risk and high-risk). Finally, these two classes were then encoded as "0" and "1", respectively.

2.2. Prediction models

We used six machine learning techniques to predict malaria outbreak risk: naive Bayes, logistic regression, KNN, decision tree, random forest, and XGBoost.

It is worth noting that, rather than a random selection, the choice of models was guided by the needs of clinical research to predict malaria outbreaks, taking into account four fundamental criteria:

Clinical applicability: We prioritized models widely validated in clinical prediction research, as they are easy for clinicians to understand and apply.

Overall performance: We evaluated model performance, focusing on adaptability to imbalanced data, generalization, and computational efficiency.

Interpretability: In line with the primary goal of machine learning to build an interpretable predictive model, we selected models that balance predictive performance with feature contribution analysis, thus avoiding the "black box" problem that would affect clinicians' confidence and the interpretation of results.

Algorithm representativeness: The selected models cover a diverse range of machine learning approaches, including probabilistic, regression-based, instance-based, tree-based, and ensemble methods, allowing for a comprehensive comparison of their performance on this malaria outbreak prediction problem.

2.3. Model evaluation

To rigorously evaluate each model's performance, we employed 10-fold cross-validation, which involves randomly splitting the training data into ten mutually exclusive folds of approximately equal size. In each iteration, nine folds were used for training and one for testing, and this process was repeated ten times to enhance statistical significance. This approach resulted in 100 performance estimates for each model, with the final performance reported as the average across all iterations, thereby ensuring robust and unbiased assessment while minimizing overfitting.

2.3.1. Hyperparameters

The optimal hyperparameter values for the machine learning models used in the malaria prediction study are shown in Table 2. To get the best values, the hyperparameters of each model were adjusted. The hyperparameter settings were derived through extensive cross-validation to balance accuracy, computational efficiency, and generalizability.

L2 regularization (Ridge) was used to optimize logistic regression with an additional strength of C=0.001. One hundred decision trees were used in random forest, with a maximum depth of 5 and 10 features considered in each split. These parameters ensure that the model captures meaningful decision patterns without overfitting. To balance the contributions of different neighbors, KNN was optimized with K=5, using uniform weighting and the Euclidean distance. The maximum depth of a decision tree was 50, and each leaf node had to contain at least 10 samples. In this way, tree growth was controlled, and unnecessary complexity was avoided. Using 100 trees, a learning rate of 0.3, and a regularization term of 0.5, the XGBoost algorithm

Table 2. Model hyperparameters.

Model	Hyperparameter	Settings	Best Value
Logistic regression	Regularization type	[Lasso, ridge]	Ridge
	Strength	[weak, strong]	Strong
		C: 1000 –0.001	C=0.001
Random forest	Number of trees	[60, 80, 100]	100
	Number of attributes considered at each split	[1, 2, …, 20]	10
	Maximum tree depth	[3, 5, …, 10]	5
K-nearest neighbors	Neighbors (K)	[1, 2, ..., 20]	5
	Distance metric	[Euclidean, Manhattan, Chebysev, Mahalanobis]	Euclidean
	Weight	[By distances, uniform]	Uniform
Decision tree	Minimum number of instances in leaves	[1, 2, ..., 10]	10
	Maximum tree depth	[10, 20, ..., 100]	50
Extreme gradient boosting	Estimator	Tree	Tree
	Number of trees	[80, 100, 120]	100
	Learning rate	[0, 0.1, ..., 1]	0.3
	Regularization type	[Lasso, ridge]	Ridge
	Limit depth of trees	[1, …, 10]	10
Naive Bayes	None	None	None

helped the model balance variance and bias. Given its intrinsic ability to adjust to various probability distributions, naive Bayes was employed without additional hyperparameter adjustments.

2.3.2. Model evaluation metrics

Classification accuracy, precision, recall, specificity, F1-score, and Matthews correlation coefficient (MCC) are the six scalar indicators used to assess prediction performance. All these metrics are calculated from a 2×2 confusion matrix, which contains the number of true positives, true negatives, false positives and false negatives.

2.4. Feature importance

Assessing feature importance is a crucial step in building effective machine learning models, particularly for disease prediction, as it identifies which attributes most strongly influence model predictions. In this study, four well-known and widely used feature ranking methods were employed with the goal of finding out which features of the data were the most informative: information gain (infGain), Gini index, *Chi*-squared, and random forest (randForest). Note that in the case of random forest used as a feature ranking approach, each tree in the forest calculates the importance of a feature based on how much it contributed to decrease the weighted impurity in the tree[28].

3. Results

3.1. Ranking the importance of features

Since each feature ranking algorithm could yield different results (rankings), it was essential to fuse the results of each algorithm using a multiple intersection approach. By examining the rankings of each algorithm, Table 3 reports the five most relevant predictors (values in bold) according to the multiple intersection method, while differences were observed in the identification of the most informative features from the sixth position onward.

The Temperature level attribute consistently received the highest ranking, highlighting its critical role in predicting malaria outbreaks. Healthcare availability and accessibility were also highly ranked, emphasizing the importance of healthcare infrastructure in epidemic control. Vector infection rates and malaria vectors were identified as important predictors, reflecting the significance of mosquito infection levels and species type. Although features such as precipitation, land use, and stagnant water ranked lower, they still contributed valuable information about environmental factors that support mosquito breeding. Employing multiple feature ranking techniques was essential for selecting the most relevant variables, thereby enhancing model accuracy and interpretability while reducing complexity and noise.

3.2. Model performance

Based on the assessment metrics, the performance of the six machine learning models is summarized in Table 4. XGBoost emerged as the most reliable model for predicting malaria outbreaks, achieving the highest values in accuracy (0.921), F1-score (0.926), precision (0.896) and MCC (0.765) along with still very high recall and specificity. Random forest also performed well, with an accuracy of 0.896, an F1-score of 0.893, a precision of 0.881, a recall of 0.902 and a specificity of 0.818, closely following XGBoost. Naive Bayes and decision tree showed moderate performance, with an accuracy of 0.769 and 0.792 or an F1-score of 0.765 and 0.785, respectively. Conversely, logistic regression and KNN had the lowest accuracy, F1-score, precision, MCC, recall and specificity values, both always below 0.700. These results demonstrate that ensemble-based learning techniques such as random forest and XGBoost outperform individual classifiers due to their ability to combine multiple decision trees, thereby reducing variance and improving generalization.

Table 3. Results of feature ranking methods*.

Predictor	InfGain	Gini	*Chi*-squared	RandForest
Mosquito species	0.001 (11)	0.001 (10)	0.086 (11)	-0.024 (13)
Mosquito breeding sites	0.005 (8)	0.002 (9)	1.99 1(7)	0.014 (8)
Precipitation	0.003 (10)	0.002 (11)	1.777 (6)	-0.004 (10)
Malaria vectors	**0.016 (3)**	**0.008 (4)**	**1.722 (5)**	**-0.030 (4)**
Presence of vegetation	0.004 (9)	0.002 (8)	1.442 (9)	-0.006 (16)
Humidity	0.000 (16)	0.000 (16)	0.864 (15)	0.006 (9)
Healthcare availability	**0.024 (2)**	**0.014 (2)**	**0.537 (4)**	**0.036 (2)**
Iot air quality	0.006 (7)	0.003 (7)	0.332 (8)	-0.046 (12)
Travel patterns	0.001 (12)	0.001 (12)	0.184 (14)	0.006 (7)
Land use	0.000 (17)	0.000 (17)	0.160 (16)	-0.024 (17)
Treatment success rate	0.012 (5)	0.006 (6)	0.123 (12)	-0.010 (5)
Stagnant water	0.000 (15)	0.000 (15)	0.085 (13)	0.028 (13)
Healthcare accessibility	**0.009 (6)**	**0.006 (5)**	**0.067 (2)**	**0.022 (6)**
Vectors infection rates	**0.016 (4)**	**0.008 (3)**	**0.055 (3)**	**-0.014 (3)**
Iot water quality	0.001 (13)	0.001 (13)	0.046 (17)	0.014 (15)
Temperature level	**0.001 (1)**	**0.001 (1)**	**0.030 (1)**	**-0.022 (2)**

InfGain: Information gain. Iot: Internet-of-Things; *the ranking of each feature is shown in brackets and the five most relevant predictors are highlighted in bold.

Table 4. Prediction performance of the machine learning models.

Model	Accuracy	F1-Score	Precision	Recall	MCC	Specificity
Naive Bayes	0.769	0.765	0.765	0.829	0.662	0.810
Logistic regression	0.611	0.472	0.632	0.609	0.536	0.634
K-nearest neighbors	0.646	0.637	0.606	0.672	0.507	0.613
Decision tree	0.792	0.785	0.781	0.781	0.683	0.782
Random forest	0.896	0.893	0.881	0.902	0.697	0.818
Extreme gradient boosting	0.921	0.926	0.896	0.826	0.765	0.810

MCC: Matthews correlation coefficient.

Table 5. Prediction performance of the machine learning models using the resampled data.

Model	Accuracy	F1-Score	Precision	Recall	MCC	Specificity
Naive Bayes	0.962	0.962	0.966	0.962	0.848	0.882
Logistic regression	0.521	0.360	0.623	0.521	0.524	0.631
K-nearest neighbors	0.646	0.645	0.668	0.646	0.516	0.646
Decision tree	0.802	0.812	0.824	0.802	0.782	0.826
Random forest	0.922	0.931	0.925	0.931	0.766	0.911
Extreme gradient boosting	0.992	0.991	0.995	0.992	0.866	0.881

MCC: Matthews correlation coefficient.

Comprehensive evaluation using metrics such as F1-score and MCC ensures a balanced assessment of model precision and recall, highlighting the importance of selecting the most suitable machine learning model based on task requirements and data characteristics.

3.3. Addressing class imbalance

Analysis of the dataset revealed a moderate imbalance: out of the 8000 samples, 4918 belong to the majority class (high outbreak risk) and 3082 to the minority class (low outbreak risk). This results in an imbalance ratio of 1.6:1, indicating that the majority class outnumbers the minority class by approximately 1.6 times. This class imbalance could negatively impact the performance of machine learning models, particularly in predicting outcomes for samples in the minority class.

It is well known that using imbalanced data to train a machine learning algorithm can favor the majority class, often resulting in higher rates of incorrect predictions for the minority class. Therefore, given the potential disparities in erroneous predictions between the majority and minority classes, we decided to address data imbalance using resampling techniques to see if any performance improvements could be achieved. Resampling was chosen because it has proven to be effective, classifier-independent, and easy to apply to any problem[29,30]. These techniques aim to modify the composition of the dataset by adjusting the number of samples in the majority and/or minority class until both classes are represented by approximately equal numbers of samples. It is important to note that the test set is neither resampled nor exposed to the machine learning model during training; that is, resampling should only be applied to the training set to avoid overly optimistic results.

Researchers argue that over-sampling is preferable to under-sampling, as the latter can discard useful data and increase the variance of the prediction models[31,32]. Taking these issues into account, this study employed SMOTE, an oversampling algorithm proposed by Chawla *et al.*[33], which involves generating minority class samples by interpolating existing samples that are very close to each other. After determining the nearest neighbors of each sample of the minority class, synthetic samples are created that point in the direction of some or all of those neighbors. Then, a certain number of samples are randomly selected from among the KNN, depending on the degree of over-sampling required.

By applying over-sampling to the imbalanced dataset, a set with an equal distribution of high and low outbreak risk was achieved (with 4918 samples in each class), thus preventing the model from being biased towards the dominant class. Furthermore, this ensures fair evaluation of high- and low-risk outbreak predictions, reduces the risk of overfitting, and improves the generalizability of the machine learning models.

Table 5 shows the performance results obtained after preprocessing the data with SMOTE. Comparing these with those in Table 4 (performance with the original data, imbalanced data), most models showed improved performance. For example, focusing on the two ensemble-based models, which correspond to those that yielded the best results with the imbalanced data, the F1-score of XGBoost increased from 0.926 to 0.991, and that of the random forest from 0.893 to 0.931. Particularly noteworthy is the significant improvement achieved by the naive Bayes method, as its F1-score increased from 0.765 to 0.962. In contrast, KNN and especially logistic regression did not benefit from oversampling, with the latter clearly performing worse after balancing the data (with an F1-score

of 0.360, an MCC of 0.524, a recall of 0.521 and a specificity of 0.631).

4. Discussion

The results of this study demonstrated the effectiveness of machine learning models in predicting the risk of malaria outbreaks based on climatic and non-climatic characteristics. The comprehensive experimental procedure included data preparation, feature selection, model training, validation, and evaluation. The dataset used in the experiments, with 8000 samples and 17 features, required extensive preprocessing to ensure its compatibility with machine learning techniques. The target variable was transformed into a binary classification problem and encoding was used to transform the categorical features into numerical representations.

A key factor in making accurate and interpretable predictions is understanding the relevance of each feature in the dataset, an aspect that is especially important in public health applications. Thus, we identified the most relevant features using four ranking algorithms—information gain, Gini index, *Chi*-squared, and random forest—combined with the multiple intersection method. This allowed us to discover that temperature level, healthcare availability and accessibility, vector infection rates, and malaria vectors were the most informative variables for predicting the risk of malaria outbreaks.

Class imbalance, which can skew model performance and lead to poor generalization, was a major problem in this study. Since the original dataset contained more samples from the "high-risk" class than from the "low-risk" class, machine learning models could potentially favor the majority class. To address this issue, the SMOTE algorithm was used, which generates synthetic samples in the minority class. By ensuring a completely even distribution of classes, this process improved the ability of most models to make accurate and fair predictions.

XGBoost achieved a maximum F1-score of 0.991, while random forest reached 0.931. The remarkable success of XGBoost and random forest is due to their ability to integrate multiple weak classifiers, which reduces variance and improves generalization. Furthermore, the highest rates across all performance evaluation metrics were obtained with the balanced data, demonstrating the importance of resampling the dataset. Compared to other studies that have also used machine learning approaches, the methodology presented in this work shows better results. For example, Okoronkwo *et al.* reported 0.756 accuracy using naive Bayes[18]. In experiments conducted by Muriithi *et al.*, the random forest emerged as the best model in predicting malaria, yielding 0.973 accuracy[34]. Similarly, Awe *et al.* obtained the highest performance using a random forest with an accuracy of 0.819[22].

There are several limitations that should be acknowledged. First, our study has focused on a single database, which could reduce generalizability. Further research across multiple databases is needed to validate the results obtained here. The present study has only experimented with traditional machine learning models, and therefore it would be interesting to investigate the performance of some deep learning methods. Furthermore, it would also be worthwhile to expand the number of techniques for addressing data imbalance, since only an over-sampling algorithm has been used here.

In conclusion, the dataset used in the experiments for this study consisted primarily of environmental features, both climatic and non-climatic, with the aim of predicting the risk of malaria outbreaks. Six machine learning models of different natures have been tested, including probabilistic, regression, lazy, and ensemble-based methods.

One of the main objectives of this study was to evaluate the potential benefits of using balanced data to make unbiased predictions favoring a majority class. To this end, the performance of the models obtained before and after over-sampling the training set has been compared. Experimental results have confirmed the importance of using balanced data when implementing a system for early prediction of malaria outbreak risk.

Overall, this study contributes to the growing field of disease prediction using artificial intelligence by demonstrating the potential of machine learning to improve surveillance of malaria and other infectious diseases, as well as outbreak management. Future research could explore larger datasets, integrate additional variables, and experiment with different machine learning approaches to further improve the accuracy of predictions. By leveraging advanced machine learning techniques, public health authorities can develop proactive malaria prevention strategies, ultimately reducing the burden of the disease in affected populations.

Conflict of interest statement

The authors declare no conflicts of interest related to this study. No financial or personal relationships could influence the work presented in this manuscript.

Acknowledgments

We express our sincere gratitude to the Reviewers and the Editor for their valuable comments and suggestions.

Funding

This study was partially funded by the Conselleria d'Innovació, Universitats, Ciència i Societat Digital–Generalitat Valenciana (grant number CAICO/2023/032) and Ministerio de Ciencia, Innovación y Universidades (grant number AIA2025-163919-C54).

Authors' contributions

N.S.E. developed conceptualization, software, and visualization. P.L.C. developed study design, methodology, and the original draft. J.S.S. developed formal analysis, supervised the study, reviewed the writing, and acquired funding. All authors read and approved the final version of the manuscript.

References

[1] Bria YP, Yeh CH, Bedingfield S. Significant symptoms and nonsymptom-related factors for malaria diagnosis in endemic regions of Indonesia. *Int J Infect Dis* 2021; **103**: 194-200.

[2] Nyasa RB, Fotabe EL, Ndip RN. Trends in malaria prevalence and risk factors associated with the disease in Nkongho-mbeng: A typical rural setting in the equatorial rainforest of the South West Region of Cameroon. *PLoS One* 2021; **16**(5): e0251380.

[3] World Health Organization. *World malaria report 2024: Addressing inequity in the global malaria response*. Geneva, Switzerland: World Health Organization; 2024.

[4] European Centre for Disease Prevention and Control. *Malaria–Annual epidemiological report for 2022*. Sweden: Stockholm; 2024.

[5] Woube M. Geographical distribution and dramatic increases in incidences of malaria: Consequences of the resettlement scheme in Gambela, SW Ethiopia. *Indian J Malariol* 1997; **34**(3): 140-163.

[6] Teklehaimanot HD, Lipsitch M, Teklehaimanot A, Schwartz J. Weather-based prediction of *Plasmodium falciparum* malaria in epidemic-prone regions of Ethiopia I. Patterns of lagged weather effects reflect biological mechanisms. *Malar J* 2004; **3**: 41.

[7] Kassa AD, Beyene BB. Climate variability and malaria transmission-Fogera district, Ethiopia, 2003–2011. *Sci J Public Health* 2014; **2**(3): 234-237.

[8] Thomson MC, Mason SJ, Phindela T, Connor SJ. Use of rainfall and sea surface temperature monitoring for malaria early warning in Botswana. *Am J Trop Med Hyg* 2005; **73**(1): 214-221.

[9] Ndiaye O, Le Hesran JY, Etard JF, Diallo A, Simodon F, Ward MN, et al. Variations climatiques et mortalité attributéee au paludisme dans la zone de Niakhar, Sénégal, de 1984 à 1996. *Cahiers Santé* 2001; **11**: 25-28.

[10]Martineau P, Behera SK, Nonaka M, Jayanthi R, Ikeda T, Minakawa N, et al. Predicting malaria outbreaks from sea surface temperature variability up to 9 months ahead in Limpopo, South Africa, using machine learning. *Front Public Health* 2022; **10**: 962377.

[11]Christiansen-Jucht C, Parham PE, Saddler A, Koella JC, Basáñez MG. Temperature during larval development and adult maintenance influences the survival of *Anopheles gambiae* s.s. *Parasit Vectors* 2014; **7**: 489.

[12]Afrane YA, Githeko AK, Yan G. The ecology of *Anopheles mosquitoes* under climate change: Case studies from the effects of deforestation in East African highlands. *Ann N Y Acad Sci* 2012; **1249**(1): 204-210.

[13]Zinszer K, Kigozi R, Charland K, Dorsey G, Kamya M, Buckeridge D. Predicting malaria in a highly endemic country using environmental and clinical data sources. *Online J Public Health Inform* 2014; **6**(1): e159.

[14]Khan O, Ajadi JO, Hossain MP. Predicting malaria outbreak in the Gambia using machine learning techniques. *PLoS One* 2024; **19**(5): e0299386.

[15]Nkiruka O, Prasad R, Onime C. Prediction of malaria incidence using climate variability and machine learning. *Inf Med Unlocked* 2021; **22**: 100508.

[16]Aheto JMK, Duah HO, Agbadi P, Nakua EK. A predictive model, and predictors of under-five child malaria prevalence in Ghana: How do LASSO, Ridge and Elastic net regression approaches compare? *Prev Med Rep* 2021; **23**: 101475.

[17]Harvey D, Valkenburg W, Amara A. Predicting malaria epidemics in Burkina Faso with machine learning. *PLoS One* 2021; **16**(6): e0253302.

[18]Okoronkwo MC, Omankwu OCB, Kanu C. The potential of machine learning techniques in predicting malaria outbreaks in Nigeria. *Caritas J Phys Life Sci* 2024; **3**(1): 9.

[19]Ayoka TO, Nnadi CO. Application of machine learning models in predicting malaria prevalence in Nigeria: An analysis of the 2015–2020 demographic and health surveys. *J Parasit Dis* 2025; doi: https://doi.org/10.1007/s12639-025-01880-6.

[20]Rahman MS, Shiddik MAB. Unraveling global malaria incidence and mortality using machine learning and artificial intelligence–driven spatial analysis. *Sci Rep* 2025; **15**(1): 28334.

[21]Monteiro KHC, da Silva Rocha É, Morais LA, Santos EG, da S Neto R, Sampaio V, et al. Integrating machine learning and spatial clustering for malaria case prediction in Brazil's Legal Amazon. *BMC Infect Dis* 2025; **25**(1): 802.

[22]Awe OO, Mwangi PN, Goudoungou SK, Esho RV, Oyejide OS. Explainable AI for enhanced accuracy in malaria diagnosis using ensemble machine learning models. *BMC Med Inform Decis Mak* 2025; **25**(1): 162.

[23]Lee YW, Choi JW, Shin EH. Machine learning model for predicting malaria using clinical information. *Comput Biol Med* 2021; **129**: 104151.

[24]Mbaye O, Ba ML, Sy A. On the efficiency of machine learning models in malaria prediction. In: Mantas J, Stoicu-Tivadar, Chronaki C, Hasman A, Weber P, Gallos P, et al. (eds.). *Public health and informatics*. Ebook: IOS Press; 2021, p. 437-441. [Online]. Avaliable from: https://ebooks.

iospress.nl/doi/10.3233/SHTI210196. [Accessed on 13 March 2026].

[25]Fleitas PE, Sarasola LB, Camprubi-Ferrer D, Muñoz J, Petrone P. Machine learning approach to identify malaria risk in travelers using real-world evidence. *Heliyon* 2024; **10**(7): e28534.

[26]Yang S, Li RY, Yan SN, Yang HY, Cao ZY, Zhang L, et al. Risk assessment of imported malaria in China: A machine learning perspective. *BMC Public Health* 2024; **24**(1): 865.

[27]Amosa RT, Abiodun IA, Biodum OA, Olatunji LM, Ifeoma UJ. Evaluating the impact of imbalanced data on malaria prediction accuracy. *Int J Res Innov Appl Sci* 2025; **10**(4): 57-65.

[28]Iranzad R, Liu X. A review of random forest-based feature selection methods for data science education and applications. *Int J Data Sci Anal* 2025; **20**: 197-211.

[29]Weiss GM. Mining with rarity: A unifying framework. *ACM SIGKDD Explor Newsl* 2004; **6**: 7-19.

[30]García S, Luengo J, Herrera F. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowl–Based Syst* 2016; **98**: 1-29.

[31]López V, Fernández A, García S, Palade V, Herrera F. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Inf Sci* 2013; **250**: 113-141.

[32]García V, Sánchez JS, Marqués AI, Florencia R, Rivera G. Understanding the apparent superiority of over-sampling through an analysis of local information for class-imbalanced data. *Expert Syst Appl* 2020; **158**: 113026.

[33]Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over–sampling technique. *J Artif Intell Res* 2002; **16**: 321-357.

[34]Muriithi D, Lumumba VW, Okongo M. A machine learning-based prediction of malaria occurrence in Kenya. *Am J Theor Appl Stat* 2024; **13**(4): 65-72.

Publisher's note

The Publisher of the *Journal* remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Edited by Lei Y, Zhang Q, Pan Y