Informatics in Medicine Unlocked 22 (2021) 100508

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: http://www.elsevier.com/locate/imu

Prediction of malaria incidence using climate variability and machine learning

Odu Nkiruka, Rajesh Prasad *, Onime Clement

Department of Computer Science, African University of Science and Technology, Abuja, Nigeria

ARTICLE INFO

Keywords:
Malaria incidence
Climate variability
XGBoost
Machine learning for healthcare and data mining

ABSTRACT

Malaria remains a serious obstacle to socio-economic development in Africa. It was estimated that about 90% of the deaths occurred in Africa, where various factors such as ecosystem and climate conditions are favorable to species of mosquitoes transmitting the malaria parasite. Some malaria epidemic prediction systems have been built to mitigate the increase of the disease outbreaks in some African countries; however, there is a need for better models with improved prediction ability based on non-seasonal variations in climatic conditions. This research proposes a machine learning-based model for the classification of malaria incidence using climate variability across six countries of Sub-Saharan Africa over a period of twenty-eight years. The work begins with a feature engineering process, which identifies the climate factors that affect the incidence of malaria, followed by the k-means clustering process for outlier detection, and then, XGBoost algorithm for classification. The results suggest that although the exact association between malaria incidence and climate variability varies from one geographic region to another, the non-seasonal changes in three climatic factors (precipitation, temperature, and surface radiation) significantly contribute to the outbreak of malaria. The proposed system was compared with other classification models, and the comparative results showed that the proposed system outperformed other classification models. The malaria incidence classification model is an early detection mechanism that helps to monitor the spread of malaria; it is a unique data-driven knowledge discovery system that will assist public health authorities in learning the effects of climate factors on health and also in developing relevant preventive and adaptive mechanisms to ensure a timelier health service in order to save lives.

1. Introduction

Malaria has persisted for a long period of time as one of the leading global health challenges, primarily prevalent in tropical and sub-tropical countries of the world. It is one of the major causes of illness and death in Sub-Saharan Africa [1]. In recent years, a lot of investments have been made to enhance malaria control and research programs, of which the World Health Organization (WHO) Global Technical Strategy (GTS) has stipulated a sum of $6.4 billion annually as the target to achieve a 90% decrease in malaria incidence and mortality rates by 2023 [2]. Despite these investments and some other eradication strategies initiated by the WHO, malaria incidence still shows some increasing trend in Sub-Saharan Africa [3].

Malaria disease is transmitted to humans through the bite of female mosquitoes (main vector) of the genus Anopheles. These vectors feed on human blood for their egg production. During the process of feeding, they transmit the plasmodium parasite [4]. As discussed by some researchers [5], the growth and maturity of this parasite mostly depend on climatic factors, which include temperature, rainfall, relative humidity, and thus, any change in climate factors would certainly exert an effect on the mosquito ecology [5]. This is why the influence of climatic and environmental variables over malaria incidence has been a predominant research focus [6]. It is well known that all the morphological (growth) processes of mosquitoes strongly depend on ambient temperature, water, and availability of stagnant water bodies. Much attention has been placed on rainfall as a principal factor for increasing the breeding sites of mosquitoes; in contrast, breeding habitats may be reduced through drought or extreme flooding [7]. In general, floods may likely increase the spread of vector-borne diseases through the creation of more breeding habitats for mosquitoes [8]. Rainfall also encourages the growth of vegetation, which provides breeding space for the vectors [9,10]. The ambient temperature may also result in faster development of the malaria parasite, which might lead to a higher incidence of malaria [11]. The available bodies of research have shown that climate

* Corresponding author.
E-mail addresses: nodu@aust.edu.ng (O. Nkiruka), rprasad@aust.edu.ng (R. Prasad), onime@aust.edu.ng (O. Clement).

https://doi.org/10.1016/j.imu.2020.100508
Received 10 June 2020; Received in revised form 23 December 2020; Accepted 24 December 2020
Available online 4 January 2021

variables have the possibility to work for or against the efforts to regulate the breeding habitat and survival of the mosquito vectors [12].

The analysis of the effects of climatic factors on malaria incidence in some West African countries was carried out using a statistical model that showed a negative correlation for temperature, rainfall, and malaria incidence in some areas and a positive association for other areas [13]. Another study was carried out to identify the association between climate factors and malaria incidence in Ethiopia using Pearson's correlation method. The results showed a positive relationship between rainfall and relative humidity in one part of the country, whereas the results from other regions showed an insignificant relationship between them [14]. It is clear from these studies that differences in geographical locations and meteorological variables may affect malaria incidence in diverse areas and in various ways. Ultimately, understanding the degree of impact of climate variability over malaria incidence is of paramount importance.

Despite all the studies that have been carried out on the prediction of malaria incidence based on climate variables, the regions of the Sub-Saharan African countries, most exposed to the malaria endemic, have not been considered or studied in detail. Therefore, with the increasing number of malaria cases in these regions, there is a dire need to address these issues and develop a cutting-edge solution that can enhance decision-making in the health sectors. This paper presents a model to analyze the effects of climate variability on six malaria endemic regions of Sub-Saharan African countries and also classifies malaria into high and low incidence cases based on climate variability. Climate variability is described as an increase or decrease in climate variables over a given period. In order to actualize leading-edge methods for the early prediction of malaria incidence, there is a need to further explore the applications of machine learning (ML) in healthcare.

ML offers the ability to extract knowledge from data to identify relevant patterns using classification. These patterns aid in medical diagnosis and decision-making. The methodology of this research work includes the feature engineering process to test the statistical significance of climate variability in malaria incidence and selects only relevant data, K-means clustering for outlier detection (if any), and Extreme Gradient Boosting (XGBoost) algorithm for classification. Feature engineering involves the use of domain knowledge to excerpt relevant features from original data by applying data mining processes [15]. K-means clustering is used to divide the dataset into different related clusters [16] to clean data and detect outliers. XGBoost classifier is a powerful ML model used to speed up the classification process and increase accuracy [17].

The key contribution of this paper is a machine learning model suitable for the binary prediction of malaria outbreaks based on non-seasonal changes in climatic factors.

The remaining parts of this paper are organized as follows. Section 2 discusses the review of related literature. Section 3 presents the system architecture and the datasets used. Section 4 presents the flowchart and algorithm of the proposed system. Section 5 discusses the results obtained from the proposed system experiments and further explains the statistical significance of the feature variables. Section 6 concludes the paper and suggests ways of extending this research paper.

2. Related works

Sub-Saharan African countries, such as Burkina Faso, Mali, Niger Republic, Nigeria, Cameroon, and Democratic Republic of Congo (DRC), are the most endemic in terms of malaria incidence [3]. Climate variability is seen as anomalies in climate variables, including precipitation, relative humidity, surface radiation, air temperature, and atmospheric pressure [18]. Research on climate-based malaria incidence prediction using the machine learning approach is not predominant, and none of the works have considered these selected countries of Sub-Saharan Africa. This paper presents a machine learning-based decision support system that classifies malaria incidence into high and low target classes based on climate variability.

An auto-regressive integrated moving average (ARIMA) and seasonal ARIMA (SARIMA) are simple stochastic time-series models that have been applied in malaria prediction research. They are used to train and then forecast future time points. Regression on the lagged values of the variable of interest is shown by the auto-regressive (AR) part. The linear combination of the error terms with a concurrent value is shown by the moving part, while the integrated (I) part specifies the result of the difference between the present and prior values. Each feature aims to make the model fit the data accurately [19]. Therefore, using time-series data with X_t, where t stands for integer index and X_t are real numbers, an ARIMA(p', q) model is obtained using Equation (1) as follows:

$$X_t - \alpha_1 X_{t-1} - \ldots - \alpha_{p'} X_{t-p'} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} \quad (1)$$

Where:

α_i Autoregressive parameters,
θ_i = Moving average parameters,
ε_t = Error terms

Adeola et al. [20] used time-series data to analyze the malaria incidence and environmental report in Nkomazi municipality, South Africa. Having shown that rainfall is the major factor that influences malaria incidence in that region, their SARIMA model was able to predict possible incidences of malaria for the next three months. Similarly, some authors [21] proposed a climate-based forecasting system to predict malaria in Andhra Pradesh, India using SARIMA model. This model forecasts malaria cases based on the current patterns of seasonal autocorrelation in the malaria case data. The predictive model shows that malaria cases are sturdily influenced by climatic factors, mainly rainfall, and temperature. Following this, there was an apparent variation in the trend of malaria discovered in the Visakhapatnam district of India. Although there was malaria transmission reported throughout the year, a higher number of malaria cases were recorded during the rainy season. In a study, authors [22] applied the ARIMA model to time-series data of malaria incidence in Afghanistan. Their model was able to identify that malaria was constantly at its peak from July to September and always reduced during January in Sub-Saharan Africa. This also suggests that high rainfall and humid temperature are possible strong factors behind the malaria outbreak and transmission. The ARIMA model is ideal for representing features of seasonal data more sophisticatedly, yet it lacks a good interpretation of covariance of weather factors. Another statistical model was proposed by Olusola et al. [23], who used dynamic regression models by combining negative binomial models with time-series models to identify the association between malaria cases and climate variables in Akure. Their findings summarized that an increase in monthly minimum temperature significantly increased the likelihood of malaria transmission, which consequently led to an increase in the number of inpatients and outpatient malaria cases. This is similar to the result of Mahdi et al. [24]. Tompkins et al. [25] proposed an early warning system of malaria using a statistical tool: VECTRI, that predicts anomalies in the rate of malaria transmission in Uganda. This early warning system involves a dynamical malaria model that predicts the entomological inoculation rate of the malaria vector. In a similar way, the stochastic lattice-based malaria (SLIM) model is another statistical malaria prediction model that is used to model variations in the abundance of *Anopheles* vector, the life cycle of *Plasmodium* parasites, and even malaria transmission using projected climate change in Kenya. This study investigated the combinatory effects of climate variation on dynamics of malaria transmissions through adjustments in the sporogonic cycle of Plasmodium brought by an increase in air temperature, and the gonotrophic cycle of Anopheles vector prompted by variations in the breeding environment caused by the acclimatory responses of vegetation under elevated (CO_2) [26]. Another statistical model proposed by Kim et al. [27] used a statistical method to identify the nonlinear and

delayed associations between malaria transmission and climate factors such as temperature and precipitation and then used SINTEX-F2 seasonal climate forecasts to predict malaria cases in Vhembe, Limpopo, and South Africa. This study only considered two climatic factors, such as temperature and precipitation, which may not be sufficient to determine the effects of climate factors on malaria cases. The results obtained from the various statistical approaches have been significant; however, an approach based on machine learning (ML) could offer better predictions based on its inherent ability to find intricate patterns and causal mechanisms from real-world data.

Applications of ML algorithms can be of benefit to the healthcare providers for the identification of effective treatments and best measures in the decision-making process. Previous studies have used machine learning techniques for malaria prediction, such as an early warning system that predicts the outbreak of malaria based on climate variables using a support vector machine (SVM). They presented a system that reads climatic information, such as temperature, relative humidity, wind speed, solar radiation, and precipitation, from free weather and geographical Application Programming Interface (APIs) and then predicts the possible incidence rate of malaria [28]. Another study in China used an ensemble of the machine learning algorithm, which is a technical framework that combines many learning algorithms to finish learning tasks and achieve good predictive performance better than what is obtained from any distinct learning algorithm. The authors basically used the stacking method to minimize generalization error through training a meta-learning algorithm to combine the predictions of various distinct primary learning algorithms to predict malaria into positive and negative using meteorological variables such as relative humidity, temperature, vapor pressure, air pressure, daily precipitation, moisture level, wind velocity, and sunshine duration [29]. Similarly, Thakur et al. [30] proposed an artificial neural network model that uses environmental variables such as vegetative index, rainfall, temperature, relative humidity, and daily report on clinical data to predict the incidence of malaria in India. Their study was aimed at identifying the exact period when malaria incidence is always at its peak and using the observed information to predict the values of malaria cases for the next years. Their study has shown that ML algorithms are successful for predicting the possible outbreak of malaria in some regions of the world. Most of these prediction models have been successful, but none of these methods could be set as the gold standard for malaria prediction based on climate variability as each method has a unique modeling norm since methods of prediction are dependent on the behavior of the study area. Table 1 presents a summary of the comparative study of the existing methods.

This research paper proposes a framework that classifies malaria incidences based on non-seasonal climate variations using the XGBoost classification model. Currently, no recent work has considered the technique used in this paper and these six malaria-endemic countries. This paper has contributed significantly by identifying hidden climatic factors that result in a high incidence of malaria in these six countries. The output of this research will help in a better understanding of the malaria transmission mechanism that can also assist in malaria intervention and eradication programs. Following this, the collected information will assist policymakers in understanding the confounding factors of malaria in a given region for the following year, then use the predictive model to predict the successive malaria incidences accurately and proactively act in response to the epidemic.

3. System architecture

This section highlights the different tools and techniques that were adopted to implement the malaria incidence classification model (MIC), as well as the metrics used for evaluating the performance of the proposed system.

Table 1
Comparative study of the existing models.

Reference (s)	Approach used	Strength	Weakness
19–20	SARIMA	Worked best with time-series data that exhibited periodic or seasonal characteristics and was able to predict the seasonal trend of malaria.	It is only suitable for a stationary or seasonal process.
21	ARIMA	The approach correctly predicted the incidence cases of malaria in Afghanistan based on averaging mechanism, often one-time step ahead of prior values using moving averages and has the capacity to detect the underlying patterns in time-series data and quantify their impact.	ARIMA model is ideal for prior data of a time series to simplify the forecast. It is not ideal for predicting non-seasonal data.
22	Binomial Model	The negative binomial model correctly identified the association between climate variable and the rate of malaria transmission.	The model only considered the association between climate variables and the rate of malaria transmission in Akure, Nigeria. This study has also used a statistical approach which is not efficient in observing the intricate patterns in data.
24	VECTRI	Uses statistical tools to predict anomalies in malaria transmission in Uganda.	The model forecasts the exact values of malaria transmission in Uganda using a statistical model, which may not be ideal for prediction in subsequent years.
25	SLIM	Ideal for capturing uncertainties and dynamisms in a dataset and provides a mechanism for the disruptions of malaria.	The model only considered the association between malaria vector and climate variable using a statistical model.
26	SINTEX-F2	The model ensured flexibility and was interpretable that describes the non-linearity and nonlinear-delayed relationship between malaria cases and weather factors.	This only considered two climate variables such as precipitation and temperature. SINTEX-F2 model may be susceptible to uncertainties, which may lead to misclassification of the effects of weather exposures to malaria cases.
27	SVM	Used partial least squares path modeling (PLS-PM) methodology to investigate the causal relationships amongst climatic variables. They further applied some machine learning algorithms to detect models that could accurately predict a malaria outbreak. SVM model gave the best prediction accuracy	SVM model relatively worked well with clear separation of a margin between target classes and is more effective in high dimensional spaces.
28	Ensemble Learning	Their results show that the performance of ensemble methodology is higher than traditional time-series models.	The study did not explore all prediction models; therefore, it may be tough to establish the most robust stacking

(*continued on next page*)

Table 1 (*continued*)

Reference (s)	Approach used	Strength	Weakness
			framework for the prediction of malaria cases.
29	ANN	The model extracted hidden complex nonlinear relationships between meteorological variables such as temperature, rainfall, relative humidity, and vegetative index using machine learning technique for prediction in Khammam district of India	Considered malaria prevalence and secondly, predicting the exact values of malaria may lead to incorrect prediction.

3.1. Study site

Six countries of the Sub-Saharan Africa region, including Burkina Faso, Mali, Niger Republic, Nigeria, Cameroon, and DRC, were selected for the study due to the endemic nature of malaria in these countries. Their geographical locations are explained as follows: Nigeria is between latitudes 4° and 14°N, and longitudes 2° and 15°E, having a population of about 206,139,589. Mali is between latitudes 10° and 25°N, and longitudes 13°W and 5°E, with a population of 20,250,833 people. Niger Republic lies between latitudes 11° and 24°N, and longitudes 0° and 16°E, with a population of 24,206,644. Cameroon is between latitudes 1° and 13°N, and longitudes 8° and 17°E, having a population of 21,917,602. Burkina Faso lies between latitudes 9° and 15°N and longitudes 6°W and 3°E, with a population of 20,321,378. and DRC lies between latitudes 6°N and 14°S, and longitudes 12° and 32°E, with a population of 84,068,091. Although the total number of incidences of malaria decreases each year worldwide, yet there has not been a significant change in these selected countries despite the investments that have been made to reduce the disease spread32. Consequently, understanding how climate variability affects these regions distinctively will help each of these countries for effective control of this disease and decision making. Fig. 1 shows the geographical map of the selected six countries.

3.1.1. Clinical data

The confirmed malaria incidence rate for a period of 28 years ranging from 1990 to 2017 for all the six selected countries, was obtained from the WHO data repository [33]. The dataset contains a normalized value of the annual confirmed incidences of malaria per 1000 population, which is the annual rate computed by dividing confirmed malaria cases by the associated population size of the country. The confirmed malaria incidences are reports of malaria incidence that have been confirmed and recorded by different hospitals and healthcare centers and then transferred to the WHO, a specialized agency of the United Nations saddled with matters relating to public health and which takes annual health reports of countries for the purpose of decision making for eliminating diseases. Fig. 2 shows the annual malaria incidence for the six selected regions for the past 28 years.

3.1.2. Climate data

The climate data was collected from the repository of the National Centre for Atmospheric Research (NCAR) [34]. The NCAR dataset contains observational data for a period of 28 years (1990–2017). The daily records obtained through earth system observations include the following variables: atmospheric pressure, surface temperature, precipitation, surface radiation, and relative humidity. This research has considered only the annual record of both climate variables and malaria incidence reports due to the availability of annual incidence reports of malaria in the six selected countries. The annual precipitation is within 1192 and 1694 mm, the annual temperature is between 25.0 and 29.5 °C, the annual relative humidity is between 40.2 °C and 45.5 °C, the annual surface radiation ranges between 220 and 240 °C, and finally, the pressure is between 99,814 and 99820pa across the six selected countries.

3.2. Dataset description

The dataset contains 28 records of the annual incidence of malaria from the WHO and annual climate data for the six countries. For each country, there were a total of 28 records covering the period of 1990–2017. The dataset has a total of 5 attributes (independent variables) that represent the climatic variables with one target (dependent) variable, which represents an increase or a decrease in malaria incidence cases. The attributes in the dataset include: precipitation, surface radiation, temperature, atmospheric pressure, and relative humidity, while malaria incidence is the target class. Table 2 presents a sample of the raw

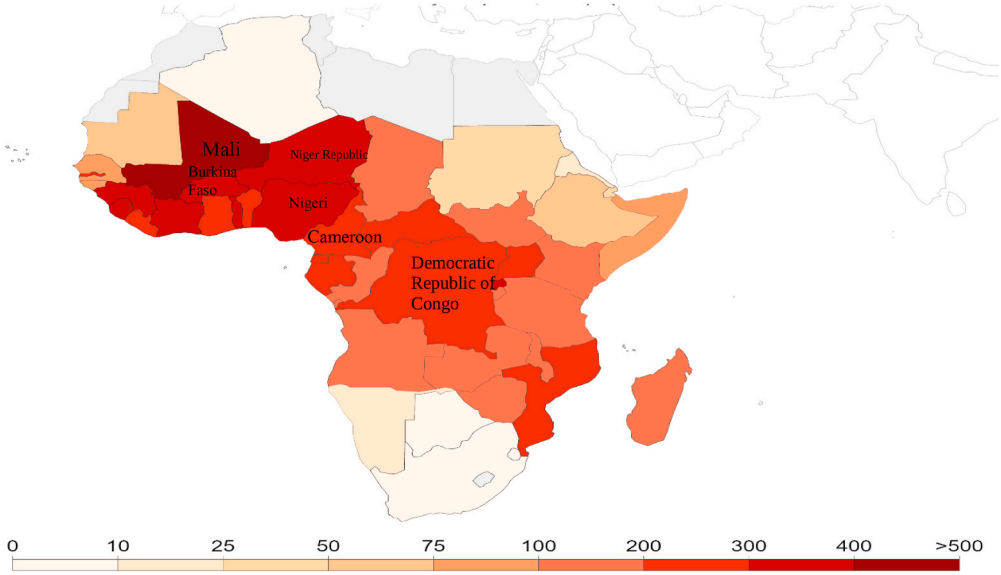

Fig. 1. Geographical map of the six selected regions and their endemicity to malaria. (source: World Health Statistics, 1990–2017).

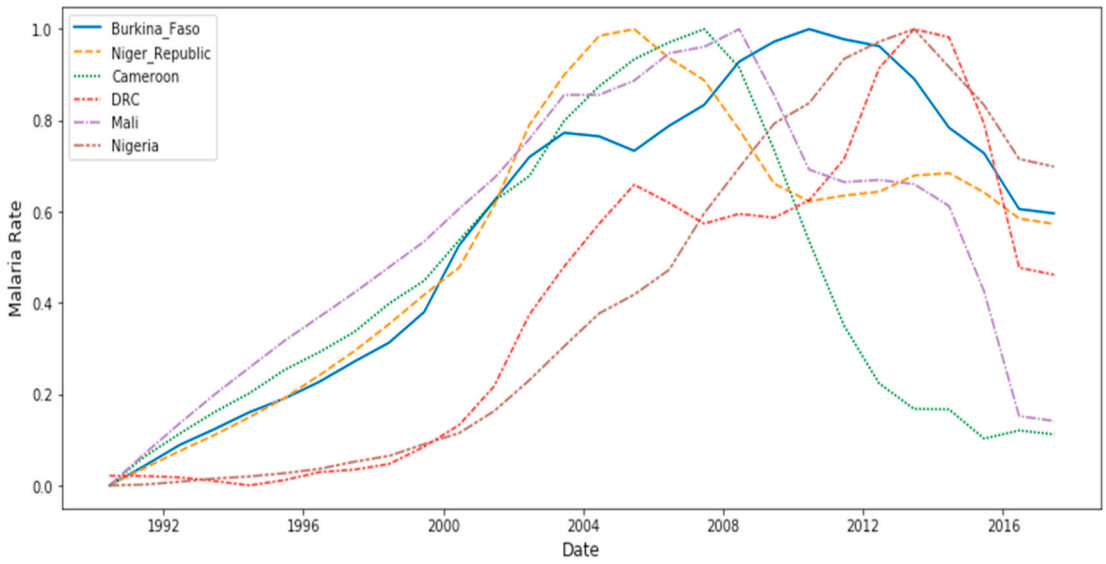

Fig. 2. Annual malaria incidence case per 1000 population for the six selected regions.
(Source: Authors).

Table 2
Sample of the dataset before pre-processing.

Date (mm/dd/yyyy)	Precipitation	R_humidity	Atm_Temp	S_radiation	Pressure	Malaria incidence
6/16/1990	0.812998	44.6705	25.5804	292.177	97208	17720.77
6/16/1991	0.980199	45.8741	25.4737	287.061	97218.8	17708.78
6/16/1992	0.851804	46.5215	25.0838	290.031	97235.7	17658.16
6/16/1993	0.83694	46.6726	25.3986	292.72	97210	17525.48
6/16/1994	1.15679	46.8185	25.4253	297.068	97232.3	17363.39
6/16/1995	0.958874	47.2742	25.5901	292.391	97201.2	17556.83
6/16/1996	0.827338	48.3952	25.5854	294.092	97144.8	17850.52

dataset before processing.

3.3. Data pre-processing

The dataset is made up of continuous variables that are heterogeneous and inconsistent due to diverse sources. In data mining principles, data quality is crucial in achieving high accuracy in the prediction of variability in malaria incidence; therefore, some pre-processing techniques [35] were applied to the dataset. We carried out an in-depth study of climate variables through the guidance of an environmentalist and a health professional to analyze the health implication of these variables to malaria transmission and occurrence; in the end, it was discovered that pressure had less significance in malaria incidence, which is in harmony with the statistical result. The dataset was normalized using minmax_scaler to unify them into the same scale. Following this, the target variable was transformed from continuous variables to discrete variables, using one of the methods of finding malaria incidence threshold as proposed by the WHO [33]. The annual mean for the past 28 years (n = 28) plus 2 multiplied by the standard deviation (SD) is shown in Equations (2)–(4) as follows:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2}$$

$$SD = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} \tag{3}$$

$$\text{Malaria incidence threshold} = \bar{x} + 2(\text{SD}) \tag{4}$$

Where:

$\bar{x}$ = *population mean*
x_i = *annual incidence report*
n = *total number of years*
SD = *Standard Deviation*

The following thresholds were obtained for each country; Burkina Faso: 13185.2, Nigeria: 215,096, DRC 2833.52, Mali: 26321.5, Cameroon: 25755.8, and Niger: 2361.83. Therefore, whenever the number of malaria incidence rises above these thresholds, it is regarded as a high incidence and vice versa. The target variable was divided into two output classes, namely: 1 and 1. At the end of the pre-processing, the datasets have no record of missing values. Table 3 presents a sample of the pre-processed data of Table 2.

4. Proposed work

This section describes in detail the flowchart and pseudocode used for the system implementation.

4.1. Feature engineering using statistical significance

The process of feature engineering in this paper involves the use of statistical correlation analysis to determine the degree of the relationship between the relative movements of two variables [36]. Pearson's correlation analysis was conducted to observe the relationship between the feature variables and target variables, which essentially investigates the strength of the association between climate variables and malaria incidence. Pearson's correlation coefficient lies between −1 and +1, where −1 shows a negative correlation, 0 indicates no correlation, whereas 1 signifies a strong positive correlation. This metric is expressed

mathematically using Equation (5). The aim of this phase is to help select only the feature variable that has a strong influence on the occurrence of malaria incidence.

$$p(x,y) = \frac{\sigma xy}{\sigma_x \sigma_y} \tag{5a}$$

where:

σ_x = *standard deviation of x*
σ_y = *standard deviation of y*
σxy = *population covariance*

4.1.1. Test for collinearity

Collinearity occurs when predictors have a linear relationship with each other. To identify predictors that have high collinearity [30], the variance inflation factor (VIF) is used to find the strength of the correlation between variables. After the examination, if the VIF is less than or equal to 1, it indicates no collinearity, but if the VIF is greater than 1, it indicates collinearity. VIF is expressed mathematically using Equations (6) and (7) as follows:

$$VIF = \frac{1}{1 - R_i^2} \tag{6}$$

Where :

i = *The predictors* $(x_1, x_2, \ldots, x_n)$

And

$$R_{adj}^2 = \left[\frac{(1 - R^2)(n-1)}{n-k-1}\right] \tag{7}$$

R_{adj}^2 = Adjusted R squared
n = total number of data samples
k = number of feature variables

4.2. K-means clustering

Clustering partitions a large number of data points into a small number of clusters. It groups the objects in such a way that objects with similar characteristics are in one group by measuring their similarity in terms of distance [16]. K-means clustering was used in this study to detect outliers and clean the dataset. Steps involved in achieving this include:

Step 1: Initializing k = 2. *k* clusters are created by assigning each input data to the nearest mean using Euclidean distance as a similarity measure, shown in Equation (8).

$$S_i^{(t)} = \left\{x_p : x_p - m_i^{(t)2} \le x_p - m_i^{(t)2} \forall j, 1 \le j \le k\right\} \tag{8}$$

Step 2: Updating the centroid for input data assigned to each cluster by recalculating the mean value or centroid, as shown in Equation (9). Steps (1) and (2) are performed iteratively until the mean value of the cluster converges.

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} x_j \tag{9}$$

Step 3: The outliers are removed by discarding the inappropriately clustered data, and a new dataset of different size is generated from this process. If the new size of the data is up to 70%, the classification process is performed; otherwise, k-means process is repeated until an appropriate data size is obtained. Apparently, about 0.012% outliers were detected and removed from the dataset at the end of the clustering processes.

4.3. Extreme Gradient Boosting (XGBoost)

Extreme Gradient boosting, popularly referred to as XGBoost, is a machine learning method that is used to solve regression and classification problems. It provides results in a prediction model, mostly in the form of trees [17]. It is scalable and efficient in memory usage and drives fast learning through parallel and distributed computing. This model is suitable for the nature of our dataset as it is important to improve the accuracy in the classes having fewer samples. The proposed MIC model is implemented using the XGBoost model. This model lets users run cross-validation in every iterative stage of the boosting process; this helps in obtaining the precise optimal number of boosting iterations in each run. Hyperparameter optimization needs to be done to achieve higher accuracy. The following hyperparameters were selected in the training of the MIC model:

i Learning_rate = 0.2; it is a step size that aids in preventing overfitting, and its value ranges between [0,1].
ii Max_depth = 6; this determines the depth of each tree that can grow during each boosting phase.
iii. n_estimators = 100, number of trees to be built.
iv. Gamma = 0.1; it regulates the splitting of a node based on the predictable decrease in loss after the split.
v. scale_pos_weight = 1; it helps in faster convergence.
vi. min_child_weight = 1; it is used to control over-fitting.
vii. Seed = 10; the random number seed can be used for parameter tuning and creating results that are reproducible.

4.4. Performance metrics

The following performance metrics were used to test the performance of the MIC model:

4.4.1. Classification accuracy

This is the ratio of the total number of accurate predictions to the total number of input samples used. This is shown in Equation (10).

$$Accuracy = \frac{number\ of\ accurate\ predictions}{total\ number\ of\ predictions} \tag{10}$$

4.4.2. Area under curve (AUC)

AUC is used for evaluating binary classification problems. AUC score provides a good summary of the performance of the receiver operator curves. A recent study shows that AUC is comparatively a better performance metric than accuracy [37]. The Receiver Operating Characteristics (ROC) curve was plotted between the false-positive rate (FPR) and true-positive rate (TPR), representing the model's performance, and then the AUC score was calculated. The sensitivity and specificity of the model were computed using Equations (11) and (12).

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{11}$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive} \tag{12}$$

4.5. System implementation

The implementation of the malaria incidence classification (MIC) model was done using Anaconda 3 that supports Python 3.6 programming language. It is open-source software that contains some packages supporting machine learning and data science applications. The design flow of the MIC is explained in Algorithm 1 and Fig. 3.

Algorithm 1. The design of the MIC model involves the following steps:

Step 1: Data preparation and Cleaning
i. *Get datasets*
ii. *Integrate datasets*
iii. *FOR each* $record_{ij}$ *of the datasets*
IF $record_{ij}$ *is NULL*
Discard record
FOR each variable
IF VIF > 1;
Discard variable;
Step 2: K-means Clustering
i. *Set number of clusters K*
ii. *Initialize the centroids = Average of all data points that belong to each cluster*
iii. *Compute the sum of squared distance between data points and centroids*
iv. *FOR each data point,*
IF the data point is closest to the cluster,
Assign data point to the nearest cluster
v. *Iterate until no change in centroids*
vi. *IF data size < 70%*
Repeat clustering process until data size >=70%
Step3: Classification and k-folds cross-validation
i. *Set number of hyperparameters, p*
ii. *Divide datasets into K-folds*
iii. *Perform parameter combination p in P*
iv. *FOR each* k_i *in k-folds*
Set fold k_i *as Test-set*
FOR fold k_j *in the remaining k-1 folds*
set k_j *as the validation set*
Train MIC model on the remaining k-2 folds
Evaluate the performance of MIC model on k_j
Calculate the average performance over k-2 to select the best parameters p
Train MIC model on the k-1 folds with the best hyper-parameters and
Get Average performance
Evaluate MIC model performance on k fold
v. *Compute the average performance over K-folds*
stop

Table 3
Sample of the dataset after pre-processing.

Precipitation	R_humidity	Air_temp	S_radiation	Pressure	Malaria incidence
−0.50512	−0.74055	−0.9594	−0.49644	0.343463	−1
−1.70873	0.486246	−1.11436	0.380656	1.149704	1
−1.22037	0.244946	−2.55947	1.720385	1.352471	1
−0.72998	0.406975	−1.26241	1.640703	2.115261	−1
1.637181	1.333713	0.10851	0.04944	−0.14897	−1
0.561079	−0.20725	−0.02379	−0.20566	−0.48692	1
0.744139	1.126933	−0.7237	0.109499	−0.55933	1

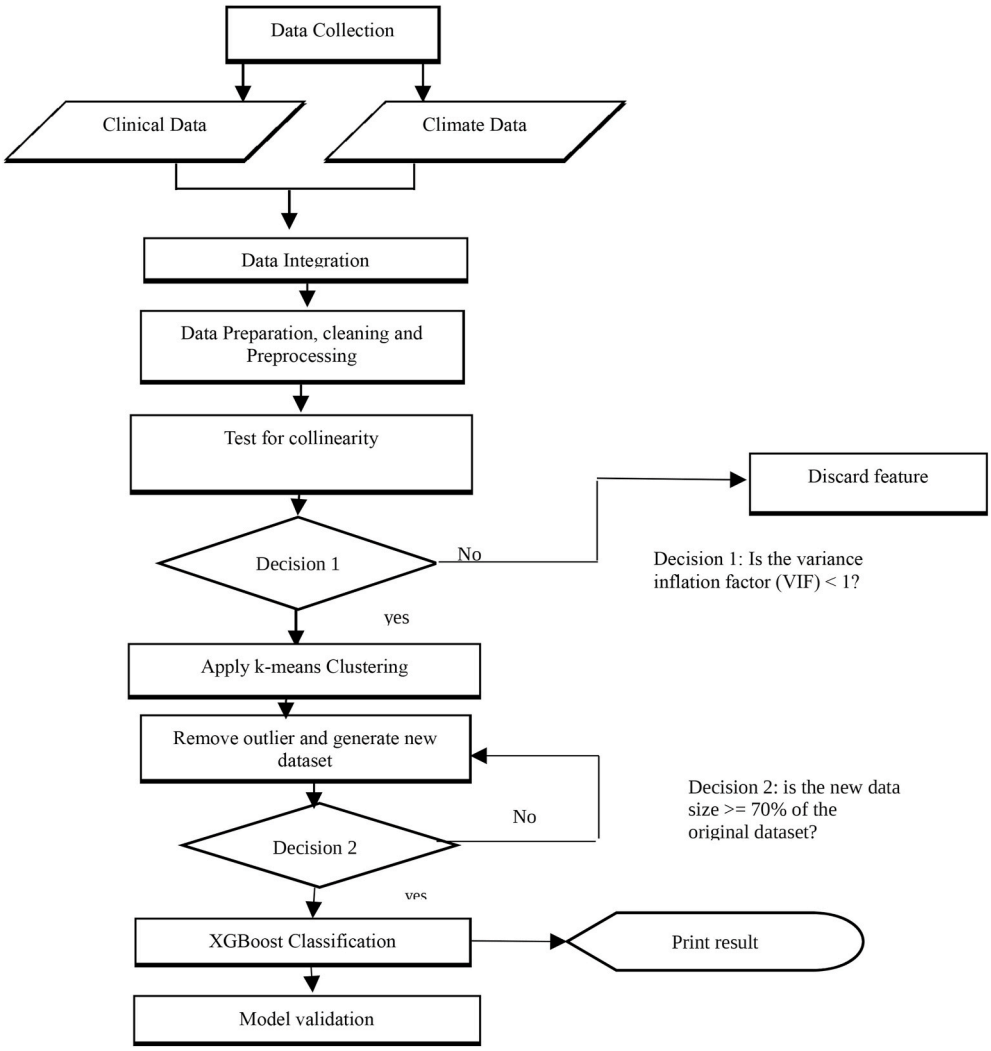

Fig. 3. Flow diagram of the Malaria incidence classification (MIC) model.

5. Results and discussion

This section presents the results obtained from the experiments and further explains the results of the comparison of the proposed MIC model with another machine learning algorithm.

5.1. Relationship between climate variability and malaria incidence

The feature engineering phase (ref. Sec. 4.1) involves the use of Pearson's correlation analysis to evaluate the relationship between the predictors and the target variable. A hypothesis test was conducted to ensure that there exists a linear relationship between the features and target variables that is strong enough to model the relationship in the sample data. If the p-value is less than the significance level ($\alpha = 0.05$), then we reject the null hypothesis, which states that there is no relationship between malaria incidence and climate variability and accepts that a significant linear relationship exists between malaria incidence and climate variability. Tables 4(a)-4(f) present the resulting correlation coefficient matrix and their corresponding p-values across the six countries. The results showed a significant variation across each country at a 95% confidence interval and $p < 0.05$.

The overall result suggests that malaria incidence has a significant positive linear relationship with air temperature and precipitation and a negative linear relationship with pressure across the six selected countries. Table 4(a) shows that a positive significant linear relationship exists between precipitation, surface radiation, relative humidity, and malaria incidence in DRC. In Table 4(b), malaria incidence in Niger shows a positive linear relationship with precipitation and surface radiation, and a negative relationship with pressure, surface radiation, and relative humidity. Similarly, in Table (4c), malaria incidence in Mali has a positive linear relationship with air temperature and precipitation, and a negative linear relationship with relative humidity, surface radiation, and pressure. Table 4(d) shows that there is a significant positive relationship between malaria incidence and precipitation, air temperature, and a negative relationship between pressure, surface radiation, and relative humidity and malaria incidence in Nigeria. Table 4(e) shows that malaria incidence has a significant positive linear relationship with precipitation, surface radiation, air temperature, and relative humidity and a negative relationship with pressure in Cameroon. Finally, Table 4 (f) shows a significant positive linear relationship between air temperature, precipitation, surface radiation, and relative humidity with malaria incidence and also a strong negative linear relationship between pressure and malaria cases in Burkina Faso.

The obtained results show that the effect of climate variability on malaria incidence rate is not homogenous but varies from one country to another. Table 5 presents a summary of the feature engineering process and shows the selected predictors based on their statistical significance to malaria incidence. The symbols "+" and "-" indicate the presence and

Table 4
Significance table.

Table 4a. DRC

X	Y	r	p-value
Precipitation	Malaria incidence	0.377095	0.047913
Pressure		−0.330831	0.085508
S-radiation		0.197373	0.314065
Air_temp		−0.573343	0.001426
R-humidity		0.053995	0.784939

Table 4b. Niger

X	Y	r	p-value
Precipitation	Malaria incidence	0.310288	0.018058
Pressure		−0.019006	0.923524
S-radiation		0.691280	0.000046
Air_temp		−0.573343	0.001426
R-humidity		−0.660484	0.0000131

Table 4c Mali

X	Y	r	p-value
Precipitation	Malaria incidence	0.211640	0.0279634
Pressure		−0.056212	0.776325
S-radiation		−0.613487	0.000517
Air_temp		0.683682	0.000061
R-humidity		−0.056212	0.776325

Table 4d. Nigeria

X	Y	r	p-value
Precipitation	Malaria incidence	0.222421	0.255285
Pressure		−0.495941	0.007276
S-radiation		−0.613487	0.000517
Air_temp		0.338675	0.0477915

Table 4e. Cameroon

X	Y	r	p-value
Precipitation	Malaria incidence	0.145549	0.459903
Pressure		−0.327888	0.088499
S-radiation		0.048337	0.807032
Air_temp		0.658249	0.000140
R-humidity		0.048247	0.807383

Table 4f. Burkina Faso

X	Y	r	p-value
Precipitation	Malaria incidence	0.399956	0.034961
Pressure		−0.595829	0.000821
S-radiation		0.015477	0.937696
Air_temp		0.694094	0.000042
R-humidity		0.041713	0.833080

absence of a predictor, respectively.

5.2. Result of MIC model

The results obtained from the feature engineering process have helped to remove irrelevant features by reducing the number of variables in the original dataset, as shown in Table 5, and this assisted in managing outliers and inconsistent data. One key advantage of feature engineering is that the few predictor variables obtained aid in reducing the dimension of the data without much loss of information, which enhances the functionality of the k-means clustering and helps in determining the number of clusters. Obtaining a good precision in machine learning model, especially in the healthcare delivery systems, is paramount and this factor must be considered before delving into implementing such systems. The dataset is split into a ratio of 70:30, where 70% (18 records) is used as the training set and 30% (8 records) as the test set. K-fold cross-validation (CV) technique at k = 5 was used during the training process. CV technique helps in reducing bias in data and thus overcomes overfitting, which may arise as a result of the quantity of the dataset [38]. The CV process is repeated *k* times, while the training set is divided into a subset of 5 distinct folds that forms a training set where each subset is used as a test set to the other 4 subsets. The *k* results are then averaged to get a single estimation. Hyperparameter optimization is first done on the training set to select the best hyperparameters using a grid search algorithm. The grid search algorithm mostly attempts all possible combinations of parameter values and then returns the combination with the maximum accuracy. For each iteration, all the possible combinations of hyperparameters are tested by fitting and scoring each combination of hyperparameters separately. In the end, the best hyperparameters are selected. The accuracy and AUC score of the MIC model are evaluated using the test set. Fig. 4 shows the ROC and AUC scores that represent the performance of the MIC model across the six countries. From Fig. 4, it can be clearly observed that the six datasets resulted in a mean AUC score of 0.97, 0.94, 0.91, 0.97, 0.94, and 0.92 for Mali, Cameroon, DRC, Nigeria, Niger, and Burkina Faso, respectively.

5.3. Comparison of results

To validate the performance of the MIC model, Table 6 presents a comparative analysis of the average accuracy score between the MIC model and other ML classification models such as Naïve Bayes, Support Vector Machine (SVM), and Logistic Regression (LR) using the same dataset based on different variations such as the original dataset, and feature engineered dataset + k-means clustering.

In Table 7, the integration of the feature engineered dataset and k-means improved the accuracy of the different algorithms modeled with the same dataset when compared to the results of Table 6. Although the results for each country vary, the proposed application of the XGBoost model still resulted in the highest accuracy measured across the six countries when compared to other classifiers. Similar to this, LR also seems to be promising as it gave close accuracy results for some datasets. Furthermore, we observed that the feature engineered dataset and k-means played a significant role in improving the accuracy of the model when compared to the results obtained using the original dataset. It is worthy to note that XGBoost worked best amongst the six different datasets, proving the MIC model to be a sound working system for classifying malaria incidence in the six selected countries.

5.4. Discussion

To validate and select the model that is the best fit for each country, we calculated the Akaike Information Criterion (AIC) and Akaike weight of the models. Akaike information criteria (AIC) is a technique used for selecting the best model during comparison [39]. AIC score is computed by maximum likelihood parameter estimation, while the Akaike weight is obtained by computing the differences in AIC value and an estimate of relative likelihood. The model with the highest Akaike weight is selected as the best classification model. We used the following number of

Table 5
Input variables (predictors) for malaria incidence classification.

Input Variable/country	Precipitation	Pressure	Surface radiation	Air Temperature	Relative Humidity
DRC	+	−	−	+	−
Mali	+	−	+	+	−
Cameroon	−	−	−	+	−
Niger Republic	+	−	+	+	+
Nigeria	−	+	+	+	−
Burkina Faso	+	+	−	+	−

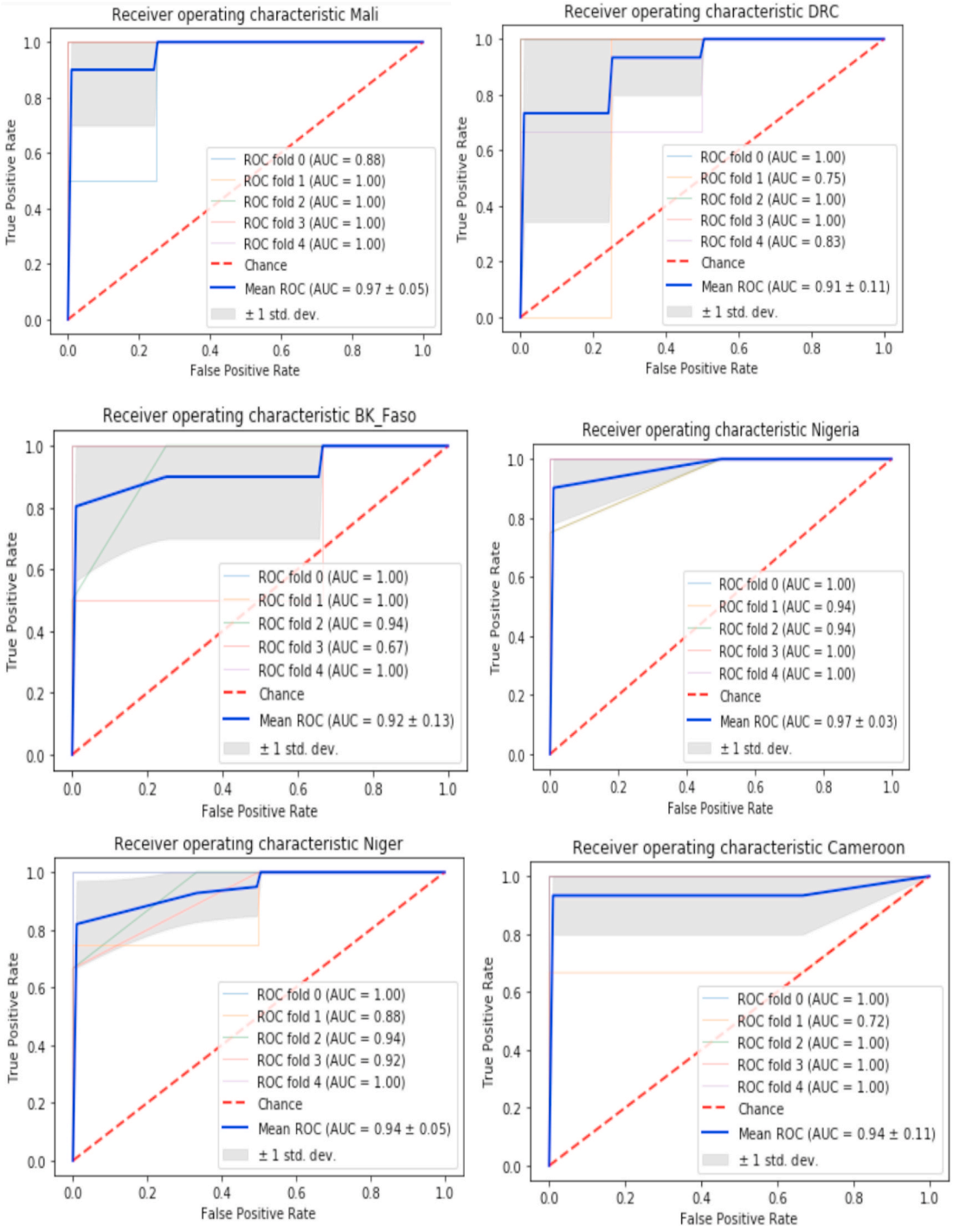

Fig. 4. ROC and AUC scores for the six countries.

Table 6
Accuracy value of original dataset without feature engineering.

Model/Country	XGBoost	SVM	Naïve Bayes	LR
Burkina Faso	0.86	0.72	0.70	0.77
Cameroon	0.86	0.68	0.72	0.70
DRC	0.81	0.72	0.71	0.76
Mali	0.82	0.70	0.74	0.75
Niger Republic	0.80	0.70	0.69	0.76
Nigeria	0.79	0.71	0.67	0.72

Table 7
Accuracy values for dataset modeled with feature engineered dataset + K-means clustering.

Model/Country	XGBoost	SVM	Naïve Bayes	LR
Burkina Faso	0.97	0.79	0.74	0.82
Cameroon	0.94	0.76	0.76	0.78
DRC	0.93	0.78	0.73	0.80
Mali	0.98	0.81	0.76	0.82
Niger Republic	0.95	0.75	0.73	0.79
Nigeria	0.98	0.74	0.71	0.81

Table 8
Akaike weight for the four fitted models.

Country/Model	MIC Model	Naïve Bayes	SVM	LR
B. Faso	0.49689	0.00752	0.03367	0.22177
Cameroon	0.49869	0.00774	0.03369	0.22037
DRC	0.49109	0.00759	0.03339	0.22602
Mali	0.49777	0.00759	0.03395	0.22097
Niger	0.49988	0.00755	0.03327	0.21982
Nigeria	0.49976	0.00739	0.03277	0.22021

parameters: MIC model: 7 parameters (ref. Sec.4.3), Naïve Bayes: 1 parameter (var_smoothingfloat is a parameter in NB, it has been set to: default = 1e-9 to signify the fraction of the principal variance of all features added to variances for calculation stability). SVM: 1 parameter (kernel = linear), and LR: 1 parameter (using a parameter: solver = 'liblinear'). Table 8 shows hypothetical results obtained from fitting four different models with the Akaike weight. It can be observed that the MIC model has the highest Akaike weight across the six different datasets; this proved that the MIC model is the best-fitted model for classifying malaria incidence in the six countries of sub-Saharan Africa.

6. Conclusion and future work

This paper presented a novel ML-based intelligent system that is capable of using real-world data to classify variations of malaria incidence based on climate variability. The results suggest that the fluctuations in malaria incidence vary with the climatic variability conditions for the six selected countries of Sub-Saharan Africa. In addition, the principal climate variable that influences malaria incidence varies from one country to another in different ways. While temperature had a strong statistical linear relationship with malaria variability in all the six sites under study, rainfall and surface radiation also showed some influence on malaria variability. With the intention to achieve a good precision for the MIC model, the feature engineering process helped to remove irrelevant features from the dataset, and k-means to clean the dataset and remove outliers, and finally, optimizing the hyperparameters of the XGBoost model also helped to improve the proposed system.

The output of this research is useful to enhance decision making towards adequate preparation for future outbreaks of malaria. This system will also aid the government of each selected country to understand the climatic factors that cause high transmission of this disease and hence regulate the environmental factors that may adversely affect the climatic conditions and thus reduce malaria incidence in their countries. It can also enhance budget making, especially when deploying eradication mechanisms such as sensitization programs and the sharing of insecticide-treated nets or malaria medicines.

This study only considered the annual data due to the unavailability of the daily reports of malaria incidence that could have helped us model the seasonal variations in the climatic factors. Secondly, improving the predictive capability of the models presented in this work requires a bigger dataset, especially for the confirmed incidences of malaria, possibly with similar or finer resolutions to the climate observations for training. Future work would involve obtaining an adequate dataset for confirmed malaria incidence, possibly as time-series data that can seasonally stratify important malaria seasons to enhance the real-time prediction by the system.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] World malaria report 2019. 2019.
[2] Brenas JH, Strecker M, Echahed R, Shaban-Nejad A. Applied graph transformation and verification with use cases in malaria surveillance. IEEE Access 2018;6:64728–41. https://doi.org/10.1109/ACCESS.2018.2878311.
[3] Africa S. Malaria ' s impact worldwide. 2019. p. 1–3.
[4] World Health Organization. Malaria surveillance, monitoring & evaluation: a reference manual. 2018.
[5] Makinde OS, Abiodun GJ, Ojo OT. Modelling of malaria incidence in Akure, Nigeria: negative binomial approach. Geojournal 2020. https://doi.org/10.1007/s10708-019-10134-x. 0123456789.
[6] Orimoloye IR, Mazinyo SP, Kalumba AM, Ekundayo OY, Nel W. Implications of climate variability and change on urban and human health: a review. Cities 2019;91:213–23. https://doi.org/10.1016/j.cities.2019.01.009. January.
[7] V Franklinos LH, Jones KE, Redding DW, Abubakar I. The effect of global change on mosquito-borne disease. Lancet Infect Dis 2019;3099(19). https://doi.org/10.1016/s1473-3099(19)30161-6.
[8] El-sappagh SH, et al. "Malaria's association with climatic variables and an epidemic early warning system using historical data from Gezira State, Sudan. Malar J 2019;13(1):1–9. https://doi.org/10.1016/j.envint.2010.03.005.
[9] Sadoine ML, Smargiassi A, Ridde V, Tusting LS, Zinszer K. The associations between malaria, interventions, and the environment: a systematic review and meta-analysis. Malar J 2018;17(1):1–11. https://doi.org/10.1186/s12936-018-2220-x.
[10] Mahendran R, Pathirana S, Sashika Piyatilake IT, Nishantha Perera SS, Weerasinghe MC. Assessment of environmental variability on malaria transmission in a malaria-endemic rural dry zone locality of Sri Lanka: the wavelet approach. PloS One 2020;15(2):1–15. https://doi.org/10.1371/journal.pone.0228540.
[11] Fischer L, Gültekin N, Kaelin MB, Fehr J, Schlagenhauf P. Rising temperature and its impact on receptivity to malaria transmission in Europe: a systematic review. Trav Med Infect Dis 2020:101815. https://doi.org/10.1016/j.tmaid.2020.101815.
[12] Pandey S, Nanda S, Vutha A, Naresh R. Modeling the impact of biolarvicides on malaria transmission. J Theor Biol 2018;454(2):396–409. https://doi.org/10.1016/j.jtbi.2018.06.001.
[13] Arab A, Jackson MC, Kongoli C. Modelling the effects of weather and climate on malaria distributions in West Africa 2014;1–9.
[14] Sena L, Deressa W, Ali A. Correlation of climate variability and malaria: a retrospective comparative study, southwest Ethiopia. Ethiop. J. Health Sci. 2015;25(2):129–38.
[15] Gárate-Escamila AK, Hajjam El Hassani A, Andrès E. Classification models for heart disease prediction using feature selection and PCA. Informatics Med. Unlocked 2020;19:100330. https://doi.org/10.1016/j.imu.2020.100330.
[16] Domingues R, Filippone M, Michiardi P, Zouaoui J. A comparative evaluation of outlier detection algorithms: experiments and analyses. Pattern Recogn 2018;74:406–21. https://doi.org/10.1016/j.patcog.2017.09.037.
[17] Ji C, Zou X, Hu Y, Liu S, Lyu L, Zheng X. XG-SF: an XGBoost classifier based on shapelet features for time series classification. Procedia Comput. Sci. 2019;147:24–8. https://doi.org/10.1016/j.procs.2019.01.179.
[18] Tompkins AM, Thomson MC. Uncertainty in malaria simulations in the highlands of Kenya: relative contributions of model parameter setting, driving climate and initial condition errors. PloS One 2018;13(9):1–27. https://doi.org/10.1371/journal.pone.0200638.
[19] Balding David J. Linear models and time-series analysis: regression. ANOVA, ARMA and GARCH; 2019.
[20] Adeola AM, et al. Predicting malaria cases using remotely sensed environmental variables in Nkomazi, South Africa. Geospat. Health 2019;14:1. https://doi.org/10.4081/gh.2019.676.
[21] Mopuri R, Kakarla SG, Mutheneni SR, Kadiri MR, Kumaraswamy S. Climate based malaria forecasting system for Andhra Pradesh, India. J Parasit Dis 2020. https://doi.org/10.1007/s12639-020-01216-6.
[22] Anwar MY, Lewnard JA, Parikh S, Pitzer VE. Time series analysis of malaria in Afghanistan: using ARIMA models to predict future trends in incidence. Malar J 2016;15(1). https://doi.org/10.1186/s12936-016-1602-1.
[23] Makinde OS, Abiodun GJ, Ojo OT. Modelling of malaria incidence in Akure, Nigeria: negative binomial approach. Geojournal 2020. https://doi.org/10.1007/s10708-019-10134-x.
[24] V Le PV, et al. Malaria epidemics in India: role of climatic condition and control measures. PloS One 2020;14(2):1–15. https://doi.org/10.1016/j.cities.2019.01.009.
[25] Tompkins AM, Colón-González FJ, Di Giuseppe F, Namanya DB. Dynamical malaria forecasts are skillful at regional and local scales in Uganda up to 4 Months ahead. GeoHealth Mar. 2019;3(3):58–66. https://doi.org/10.1029/2018GH000157.
[26] V Le PV, Id PK, Ruiz MO, Mbogo C, Muturi J. Predicting the direct and indirect impacts of climate change on malaria in coastal Kenya. 2019. p. 1–18.
[27] Kim Y, et al. Malaria predictions based on seasonal climate forecasts in South Africa: a time series distributed lag nonlinear model. Sci Rep 2019;9(1):1–10. https://doi.org/10.1038/s41598-019-53838-3.
[28] B. Modu, N. Polovina, Y. Lan, S. Konur, and A. T. Asyhari, "Applied sciences towards a predictive analytics-based intelligent malaria outbreak warning system †," pp. 1–20, doi: 10.3390/app7080836.
[29] Wang M, et al. A novel model for malaria prediction based on ensemble algorithms. PloS One Dec. 2019;14(12):e0226910. https://doi.org/10.1371/journal.pone.0226910 [Online]. Available.

[30] Thakur S, Dharavath R. Artificial neural network based prediction of malaria abundances using big data: a knowledge capturing approach. Clin. Epidemiol. Glob. Heal. 2019;7(1):121–6. https://doi.org/10.1016/j.cegh.2018.03.001.
[32] WHO. Malaria eradication: benefits, future scenarios and feasibility. Executive summary, WHO Strategic Advisory Group on Malaria Eradication. 2019. p. 20 [Online]. Available, https://www.who.int/publications-detail/strategic-advisory-group-malaria-eradication-executive-summary.
[33] Roser M, Ritchie H. Malaria. *Our World Data*; 2020.
[34] National center for atmospheric research. 2019. https://ncar.ucar.edu/what-we-offer/data-services. Data.
[35] Alexandropoulos SAN, Kotsiantis SB, Vrahatis MN. Data preprocessing in predictive data mining, vol. 34; January. 2019.
[36] Zhao H, Lai Z, Leung H, Zhang X. Feature learning and understanding. 2020.
[37] Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng 2005;17(3):299–310. https://doi.org/10.1109/TKDE.2005.50.
[38] Wong T, Yeh P. Reliable accuracy estimates from K-fold cross validation. IEEE Trans Knowl Data Eng Aug. 2020;32(8):1586–94. https://doi.org/10.1109/TKDE.2019.2912815.
[39] Wagenmakers EJ, Farrell S. AIC model selection using Akaike weights. Psychon Bull Rev 2004;11(1):192–6. https://doi.org/10.3758/BF03206482.