International Journal of Environmental Health Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cije20

Determining suitable machine learning classifier technique for prediction of malaria incidents attributed to climate of Odisha

Pallavi Mohapatra, N. K. Tripathi, Indrajit Pal & Sangam Shrestha

To cite this article: Pallavi Mohapatra, N. K. Tripathi, Indrajit Pal & Sangam Shrestha (2021): Determining suitable machine learning classifier technique for prediction of malaria incidents attributed to climate of Odisha, International Journal of Environmental Health Research, DOI: 10.1080/09603123.2021.1905782

To link to this article: https://doi.org/10.1080/09603123.2021.1905782

Published online: 26 Mar 2021.

INTERNATIONAL JOURNAL OF ENVIRONMENTAL HEALTH RESEARCH
https://doi.org/10.1080/09603123.2021.1905782

Determining suitable machine learning classifier technique for prediction of malaria incidents attributed to climate of Odisha

Pallavi Mohapatra[a], N. K. Tripathi[a], Indrajit Pal[b] and Sangam Shrestha[c]

[a]Remote Sensing and Geographic Information System, Asian Institute of Technology, Pathum Thani, Thailand; [b]Disaster Preparedness Mitigation and Management, Asian Institute of Technology, Pathum Thani, Thailand; [c]Water Engineering and Management, Asian Institute of Technology, Pathum Thani, Thailand

ABSTRACT

This study investigated the influence of climate factors on malaria incidence in the Sundargarh district, Odisha, India. The WEKA machine learning tool was used with two classifier techniques, Multi-Layer Perceptron (MLP) and J48, with three test options, 10-fold cross-validation, percentile split, and supplied test. A comparative analysis was carried out to ascertain the superior model among malaria prediction accuracy techniques in varying climate contexts. The results suggested that J48 had exhibited better skill than MLP with the 10-fold cross-validation method over the percentile split and supplied test options. J48 demonstrated less error (RMSE = 0.6), better kappa = 0.63, and higher accuracy = 0.71), suggesting it as most suitable model. Seasonal variation of temperature and humidity had a better association with malaria incidents than rainfall, and the performance was better during the monsoon and post-monsoon when the incidents are at the peak.

ARTICLE HISTORY
Received 15 March 2021
Accepted 15 March 2021

KEYWORDS
Machine learning; malaria prediction; J48 decision tree; multilayer perceptron

Introduction

Malaria remains one of the perennial public health concerns in many parts of the world, even with the efforts put in place by the World Health Organization (WHO) and other international and national bodies to curb it (World Health Organization [WHO] 2018). Malaria is characterized by seasonal transmission and distribution of vectors and is influenced by seasonal climatic variations (Kovats et al. 2003; Bomblies 2012). Both vectors and parasites tend to be sensitive to atmospheric temperature and moisture changes (Githeko et al. 2000). The distribution of malaria is limited by the climate tolerance of the mosquito vectors and the biological restrictions that limit the incubation and the survival of the infective agent in the vector population (Leal Filho et al. 2018). The examination of how climate conditions could affect malaria spreading can be approached by closely monitoring various aspects of climate change and the surrounding environment. Van Lieshout et al. (2004) examined the Spatio-temporal effects of climate change on malaria and concluded that significant changes in the temperature and rainfall patterns could increase malaria spreading. Malaria prevalence depends on the parasite *Plasmodium* and population dynamics of *Anopheles* mosquito (Holt et al. 2002; Kim et al. 2012). The development and the survival rates of both *Plasmodium* parasites and the *Anopheles* mosquitoes are dependent on weather (Parham and Michael 2010). As Kakmeni et al. (2018) explain, the temperature is key to these parasites' persistence. Current evidence suggests that inter-decadal and inter-annual variability of the climate directly affects some critical vector-borne diseases (Kovats et al. 2003; Pramanik et al. 2020).

CONTACT Pallavi Mohapatra mohapatra.pallavi@gmail.com Department of Remote Sensing and GISAsian Institute of Technology, 58 Moo 9, Paholyothin Road, Khlong Luang,Pathumthani-12120, Thailand

The optimal temperature range for vector-borne diseases' transmission remains 14–18°C at the lower-end and 35–40°C at the upper-end (World Health Organization[WHO] 2018). However, at a temperature ranging between 30°C and 32°C, the breeding and transmission rate increases substantially (Githeko et al. 2000; Dhiman et al. 2010; Parham and Michael 2010; Ngarakana-Gwasira et al. 2016). The best condition for developing malaria parasites is between 20°C and 30°C for temperature and 60% for humidity (Mishra 2003; Dhiman et al. 2008; Ngarakana-Gwasira et al. 2016). Segun et al. (2020) further described that 60–90% of relative humidity could enhance the breeding and multiplication of the Plasmodium parasite. Extended monsoon rainfall in the Indian region is significantly related to malaria breeding (Gupta 1996). However, rainfall in the range of 15–17 mm daily is ideal for malaria spread (Ngarakana-Gwasira et al. 2016), while excess rain may destroy breeding sites (Devi and Jauhari 2006).

Odisha, an eastern coastal state in India, has the maximum number of malaria incidents and casualities since 2014, compared to other states (provinces) of India, as per the National Vector Borne Disease Control Programme (NVBDCP n.d.), malaria situation report. The current research performs an analysis using advanced machine learning approaches to determine how different climate conditions are related to the transmissions of malaria and the potential of accurately predicting malaria incidence for the Sundargarh district in Odisha, India. Sundargarh district has the second-highest malaria incidents in the state, followed by the Rayagada district. A block-wise cumulative incidence map for the district is presented in Figure 1 and 2, which suggests the eastern region is severely affected by malaria, in particular. The study has evaluated the efficiency of the machine learning algorithms to identify a suitable method that will assit in predicting malaria incidents in future. Further, the findings would also encourage better utilization of climate forecasts to predict potential malaria risk.

Malaria as a Public Health Concern

There were 219 million malaria cases globally in 2017 (World Health Organization [WHO] 2018), and an estimated 228 million malaria cases occurred worldwide in 2018 (World Health Organization

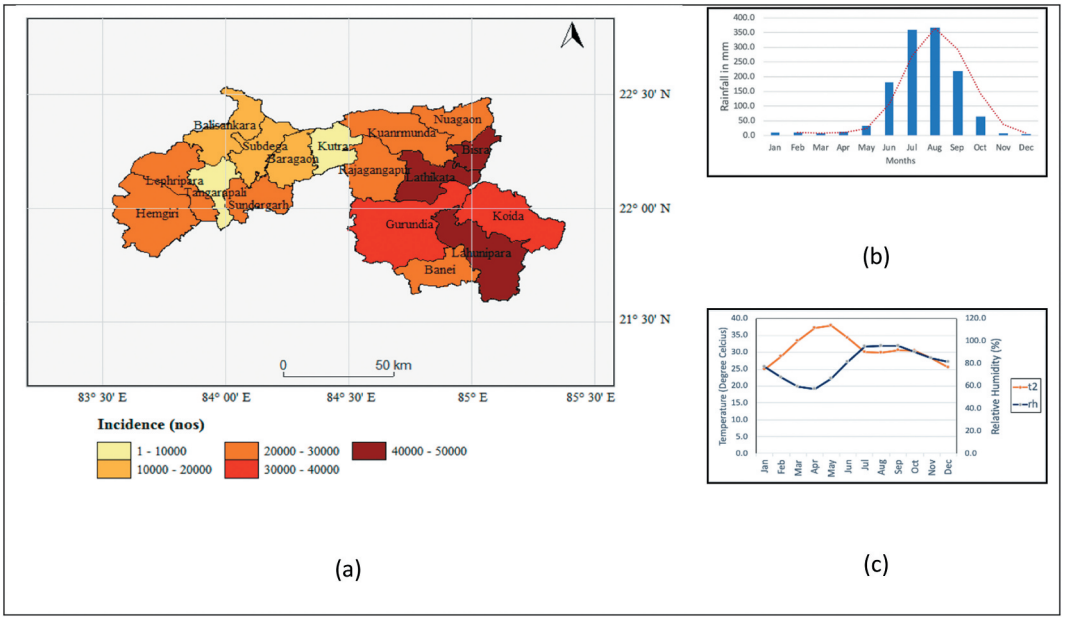

Figure 1. Cumulative Malaria Incidence Map (a) for the period 2002 to 2017; (b) monthly average rainfall; and (c) Average monthly temperature (T2) and relative humidity (RH) (Data Source: Directorate of Public Health, Odisha).

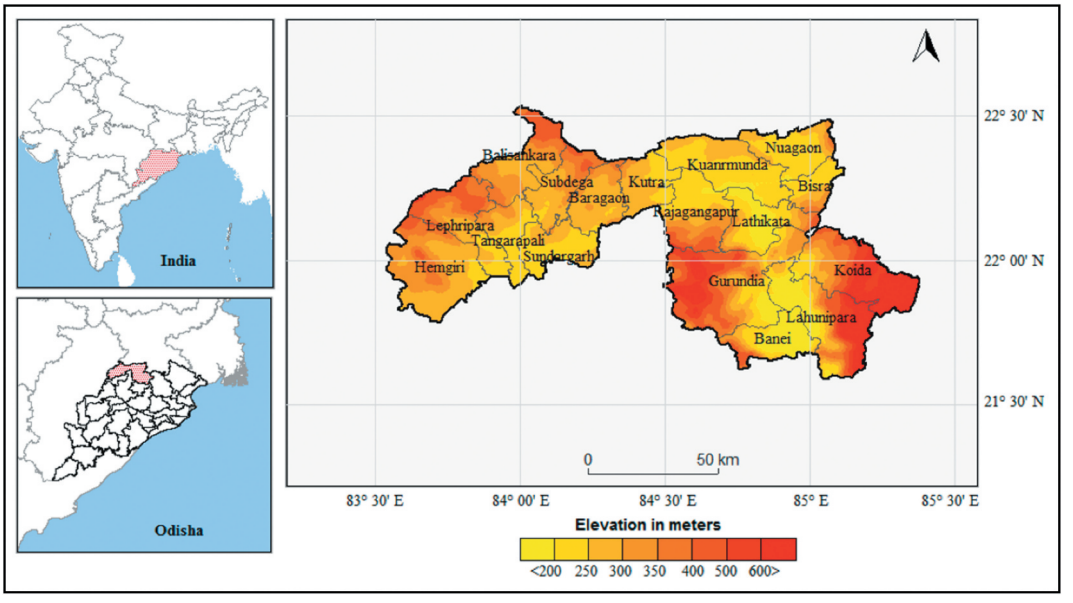

Figure 2. Study area with elevation for Sundargarh district in Odisha, India.

[WHO] 2019). The burden was most substantial in the African region, where an estimated 93% of all malaria deaths occurred, and in children aged under 5 years, who accounted for 61% of all deaths (World Health Organization [WHO] 2018). Almost 85% of all malaria cases globally were in 19 countries, including India and 18 African countries. In India, seven states accounted for 90% of the estimated cases in 2018, counting to 5.7 million cases (World Health Organization [WHO] 2019). Malaria is prevalent in eastern, central, and north-eastern states, especially in ethnic groups and tribal populations. Inequality and poverty in this area play a crucial role in the spreading as well (Kannan 2017; Mahakur and Nayak 2019) and who habitually reside in remote areas with a complex topography and dense forest with limited access to basic facilities (Sundararajan et al. 2013).

Influence of Climate on Malaria

Several researchers used statistical methods to investigate the association between climatic factors and malaria incidents. For example, the multiple polynomial regression to model malaria incidents in India (Chatterjee and Sarkar 2009), semi-parametric Poisson distribution methods to model the influence of temperature and rainfall on malaria incidence in Zambia (Shimaponda-Mataa et al. 2017), distributed non-linear lag model to associate malaria to meteorological factors in China (Guo et al. 2015), hierarchical Bayesian framework to model effects of weather and climate on malaria distributions in West Africa (Arab et al. 2014) and the time series regression models (Imai et al. 2015; Rejeki et al. 2018) are few of those researches. All the models have shown reasonable skills over the respective regions. Neter et al. (1996) recommend using Multiple Linear Regression (MLR) to analyze data because this model can determine the relative influence of one or more predictor variables. Other advanced computational models include Artificial Neural Network (ANN) models, which are relatively simple to interpret (Basheer and Hajmeer 2000) and, as such, require less formal training. They also can, implicitly, detect complex non-linear relationships between the set of variables being investigated. Yao et al. (1999) demonstrated that using a neural network for data analysis could detect all possible predictors' possible interactions.

Influence of non-climatic factors on Malaria

Apart from climate, non-climatic factors also play a critical role in the spreading of malaria. Environmental factors, including land cover, topographical variations, and human activities like farming with highly cultivated areas, have perpetuated some mosquitoes' survival (Macherera & Chimbari 2016). Besides, rapid urbanization with decreased sanitary conditions, mass population displacement promote vector breeding in some instances (Mouchet et al. 1998, Williams et al. 2003, Tatem et al. 2004, Hay et al. 2005). Land reforms through large construction projects provided the breeding sites for malaria vectors, leading to malaria epidemics in Rajasthan, India (Lingala 2017). Socio-economic risk factors have also contributed to the spread of malaria in different regions. Poverty increases the risk of malaria spreading (Gallup et al. 2001). People with lower income, living conditions with house type, distance to health facilities, availability, and use of mosquito nets were positively associated with malaria in the Indian region (Yadav et al. 2014) and the African region (Lowassa et al. 2012) for malaria spreading in northern Tanzania. The above studies indicate that malaria outbreaks are multi-factorial and depend on environmental factors' changing conditions.

Research Outcome

This research acknowledges the existing problem of the malaria epidemic in the region and the influence of increased frequency of extreme climate events contributing to the escalation of malaria spreading. This research aims to classify malaria incident data and its association with climate variables like the rainfall, daily maximum temperature and relative humidity. With the WEKA tool's help the study would identify a suitable machine learning classifier technique among two methods, MLP, a Neural netwok-based approch and J48, a decision tree-based approch with three sets of supplied test options. The authors would classify the dataset and then analyze the algorithm best suited for predicting malaria with a better accuracy, skill and lower error. The expected research outcomes were to i) identification of the superior classifier method and test options with consistent performance over the district and the 17 blocks, and ii) evaluation of the seasonal influence of weather on the malaria incidence.

Materials and methods

Study area

The National Vector Borne Disease Control Programme statistics revealed that Odisha, a coastal province in India, records the maximum number of casualties due to vector-borne diseases, especially malaria (NVBDCP n.d.). This research targeted the Sundargarh district, which records one of the highest numbers of malaria incidents in the state. The geographical location of Odisha made it susceptible to increased occurrences of climate extremes and adversely affecting human

Table 1. Data source and the attributes.

Type of Data	Data Source	Period	Spatial Scale	Temporal scale
Malaria Incidents	The Directorate of Public Health, Odisha	2002 – 2017	Block	Monthly
Rainfall	Special Relief Commissioner, Odisha	2002–2017	Block/Station	Daily
Surface Max. Temperature	ECMWF Reanalysis land data (ERA5-Land)	2002–2017	Gridded. 0.1°x0.1°; Native resolution is 9 km	Daily
Relative Humidity	ECMWF Reanalysis land Data (ERA5-Land)	2002–2017	Gridded. 0.1°x0.1°; Native resolution is 9 km	Daily

health due to the environmental changes. Sundargarh district forms the north-western part of Odisha state and is the second-largest district in the state, accounting for 6.23% of the total area. The geographical location of the district is 9712 square km. The district spreads from 21°36′N to 22°32′N and 83°32′E to 85°22′E (Directorate of Census Operations, Odisha [DCO] 2011).

Topography

The district exhibits an ideal ecological condition for malaria transmission topographically with its undulating uplands intersected by forested hills and widely diversified tracts of mountains. The areas covered by western blocks are long undulating tracts of about 700 ft. (213 mt.) above the sea level, dotted with hill-ranges and isolated peaks of considerable height. Simultaneously, the far eastern and southern-central blocks are mostly an isolated hilly tract with an average elevation of about 800 ft. (244 mt.) above sea level.

Climate of Sundargarh

The area received rainfall between June and September and is described as a tropical humid climate region from the southwest monsoon. The average annual temperature ranges between 22°C and 27° C, and the average annual rainfall ranges between 1600 and 2000 mm. The weather seasons are hot and dry summer from April to mid-June, monsoon from mid-June to September, autumn from October to November, winter from December to January, and spring from February to March. The maximum temperature during summer rises to 40–45°C, and the minimum temperature during winter falls to 5–10°C.

Data used

The data used for the study include rainfall (RF), relative humidity (RH), and surface (2-meter height from ground), maximum temperature ($T2_{max}$), and malaria incidents. While the climate parameters are treated as independent variables, the malaria incident records are considered the dependent variable. For this analysis, historical meteorological rainfall data from 17 blocks for the district are accessed from the Odisha Government portal of special relief commissioner (http://srcodisha.nic.in/rain_fall.php), which is a publicly accessible portal. Surface maximum temperature and relative humidity data with a horizontal resolution of 0.1-degree was obtained from the Copernicus Climate data store (CDS) of the European Center for Medium-Range Weather Forecast (ECMWF). The most recent ECMWF Reanalysis (ERA5-Land) is a reanalysis of the global atmosphere covering the data-rich period since 1981 and continuing in real-time. More details about the dataset can be found from the Copernicus climate data store (https://cds.climate.copernicus.eu/cdsapp#!/dataset/10.24381/cds.e2161bac?tab=overview). The data was taken at daily temporal time scales for both relative humidity and surface temperature. Many studies have demonstrated the use of gridded reanalysis data in the absence of actual ground observation (Bengtsson 2004; Jolivet et al. 2011; Parker 2016; Marcos et al. 2019) as it is the closest possible representation of the actual observations. Monthly malaria incident datasets at block level were collected from the Directorate of Public Health Services, Government of Odisha. For consistency in the analysis, all data were collected for the identical period of 2002 to 2017. Table 1 summarizes the data used in the study.

Data preparation

The data collected for the analysis were at different temporal scales and required to be brought to a standard spatial and temporal scale. As the malaria incidents are the parameter to be predicted and were available at a monthly scale, the climate data (which are at daily time scales) were statistically averaged over the month. The ERA5-Land data for maximum temperature and relative humidity were extrapolated to produce average spatial data for the blocks. The percentile (p = 25, p = 50, p = 75, and

p = 95) were computed for each sample to define the variables' spread. The historical data was converted from numerical to nominal (Low, Medium, High, and Very High) ranges for the analysis.

Weka machine learning tool

Waikato Environment for Knowledge Analysis (WEKA) is a collection of machine learning algorithms that accurately perform data mining tasks (Weka 1994). WEKA contains tools that facilitate data preparation, regression, classification, association rules mining, clustering, and visualization. Through its machine learning platform, WEKA enables the algorithm to learn about data as samples and with or without any other explicit programs (Witten and Frank 2002; Hornick 2009). More detail about the tool is available at https://www.cs.waikato.ac.nz/~ml/weka/. Multilayer Perceptron (MLP)) and J48 classifier techniques in the Weka tool are commonly used to predict malaria incidents (Sharma et al. 2015; Bui et al. 2019; Olayinka and Chiemeke 2019). Researchers around the globe also used it for prediction of dengue (Shakil et al. 2015; Guo et al. 2017; Atulbhai 2017; Mello-Román et al. 2019) and other public health issues such as Cholera (Leo et al. 2019), diabetes (Al Jarullah 2011; Zia and Khan 2017; Mahmud et al. 2018), heart diseases (Dangare and Apte 2012; Sabarinathan and Sugumaran 2014). Both these methods are therefore used in the current study.

Multilayer perceptron

A Multilayer Perceptron (MLP) is a class of feed-forward artificial neural networks (Korting 2006). It constitutes at least three layers of nodes, a hidden layer, an input layer, and an output layer. Each of these nodes, except the input nodes, is a neuron that uses a non-linear activation function. For a long time, ANN has been a robust perceptive classifier for tasks not just in medical diagnosis but also for early detection of diseases (Nicholson n.d). MLP uses a supervised learning technique to propagate the network (Korting 2006); it modifies the standard linear perceptron. As such, it can distinguish data that is not separable. A perceptron produces a single output based on several real-valued inputs by forming a linear combination using its input weights. Which can be represented in the following form:

$$y = \varphi\left(\sum_{i=1}^{n} w_i x_i + b\right) = \varphi(w^T x + b\Big) \tag{1}$$

where $\boldsymbol{w}$ denotes the vector of weights, $\boldsymbol{x}$ is the vector of inputs, $\boldsymbol{b}$ is the bias, and *is the non-linear activation function.*

*MLP is composed of an input layer to receive the signal, an output layer that decides or predict the input, and in between those two, an arbitrary number of hidden layers that are the actual computational engine of the MLP. They train on a set of input-output pairs and learn to model the correlation (or dependencies) between those inputs and outputs (*Nicholson n.d.). Training involves adjusting the parameters, or the weights and biases, of the model to minimize error. Back-propagation is used to make those weight and bias adjustments relative to the error, and the error itself can be measured in a variety of ways.

J48 classifier model – A decision tree based method

J48 in WEKA is the implementation of the C4.5 decision tree (Korting 2006; Quinlan 2014). Dangare and Apte (2012) defined J48 classification as building models of classes from records that contain class labels. A decision tree algorithm is used to find how the attribute-vector is likely to behave for an array of instances. The algorithm generates rules that would predict the targeted variables and accounts for any missing values present in the model and the output. Some algorithms perform classification recursively until each leaf has been deemed pure (Korting 2006). In other words, the classification of data would be as perfect as possible. The objective of the J48

classification is to reduce the impurity in data as much as possible. A subset of data is considered pure if all instances belong to the same class. The heuristic is to choose the attribute with the maximum Information Gain or Gain Ratio based on information theory. Entropy is a measure of the uncertainty associated with a random variable. We choose the attribute with the highest gain to split the current tree.

Assuming the attributes are categorical, a tree is constructed in a top-down recursive manner. At the start, all the training samples are at the root, and samples are partitioned recursively based on selected attributes. Attributes are selected based on an impurity function (e.g., information gain). This process uses the 'Entropy,' i.e., measuring the data's disorder (Korting 2006; Quinlan 2014; Kaur and Chhabra 2014). The Entropy of $\overrightarrow{\mathbf{E}}$ is calculated as:

$$\mathbf{Entropy}\left(\overrightarrow{\mathbf{E}}\right) = -\sum_{\mathbf{j}=1}^{\mathbf{n}} \frac{\left|\mathbf{E_j}\right|}{\left|\overrightarrow{\mathbf{E}}\right|} \log \frac{\left|\mathbf{E_j}\right|}{\left|\overrightarrow{\mathbf{E}}\right|} \tag{2}$$

iterating over all possible values of $\overrightarrow{\mathbf{E}}$. The conditional Entropy is

$$\mathbf{Entropy}\left(\mathbf{j}|\overrightarrow{\mathbf{E}}\right) = \frac{\left|\mathbf{E_j}\right|}{\left|\overrightarrow{\mathbf{E}}\right|} \log \frac{\left|\mathbf{E_j}\right|}{\left|\overrightarrow{\mathbf{E}}\right|} \tag{3}$$

and finally, the gain is

$$\mathbf{gain}\left(\overrightarrow{\mathbf{E}}, \mathbf{j}\right) = \mathbf{entropy}\left(\overrightarrow{\mathbf{E}} - \mathbf{Entropy}\left(\mathbf{j}|\overrightarrow{\mathbf{E}}\right)\right) \tag{4}$$

The aim is to maximize the gain, dividing by overall Entropy due to split argument by value j.

Predictive modeling using MLP and J48

The climate datasets were reprocessed to a monthly scale as the prediction of malaria incidents was intended to be carried out at a monthly time scale. The numerical monthly malaria incident and climate data were then transformed to the nominal range before they were fed to the MLP and J48 classifier models. The datasets were split into two sets; the first set for training of the model and the second set for testing or the prediction. The test options used include; (a) 10-fold cross-validation method, where all samples were divided as ten equal sets, from which one set is used for testing, and the remaining nine sets for the model's training. (b) Percentage split method, where data is distributed as a percent of the total number of samples with 34% data for testing, and 66% data are used for training. (c) A supplied test set enables users to decide on the distribution of the samples for training and prediction.

Performance indicators

With the classifiers, we investigate how good both the models are, and this is done by examining the number of correctly classified instances to the number of incorrectly classified cases from the supplied datasets. The machine learning analysis methods' performance was evaluated through different indicators that were inbuilt in the tool. These include the Root Mean Square Error (RMSE), the accuracy, the kappa, and the Receiver Operating Characteristics (ROC) values. This section provides a brief about each of these indicators and their significance. A confusion matrix offers a simplified structure of the representation of the observed and expected samples to segregate the classifications into four classes: True Positives (***a***), False Positives (***b***), False Negatives (***c***) and True Negatives (***d***) and Total Number of samples (***N***).

The observed agreement is the frequency with which the two variants (observed and expected) agreed. The observed and expected agreement can be determined as (Viera and Garrett 2005):

$$\text{Observed agreement}(\mathrm{P_o}) = \frac{\mathbf{a}+\mathbf{d}}{\mathbf{N}} \tag{5}$$

$$expected\ agreement\ (p_e) = expected\ (a)\ +\ expected\ (d) \tag{6}$$

Where expected$(\mathbf{a}) = \frac{\mathbf{a+b}}{\mathbf{N}} * \frac{\mathbf{a+c}}{\mathbf{N}}$ **== and** $expected(d) = \frac{b+d}{N} * \frac{c+d}{N}$

Accuracy

The percentage of correctly classified instances is often called accuracy. The basic formula for calculation of prediction accuracy can be described as (referring to the confusion matrix for observed agreement)

$$accuracy = \frac{a+d}{N} \tag{7}$$

Where $a = TruePositives$ and d $= CorrectNegatives$.

Kappa coefficient

Kappa is the measurement of the inter-rater reliability, representing the extent to which the data collected in the study are correct representations of the variables measured (McHugh 2012). Kappa can be represented as:

$$Kappa = \frac{p_0 - p_e}{1 - p_e} \tag{8}$$

Where $\mathbf{P_o}$ = Observed agreement; $\boldsymbol{P_e}$ = *Expected agreement*

Kappa coefficients are interpreted using the guidelines outlined by Landis and Koch (1977), where the strength of the kappa coefficients is interpreted in the following manner: 0.01–0.20 slight; 0.21–0.40 fair; 0.41–0.60 moderate; 0.61–0.80 substantial; 0.81–1.00 almost perfect. A negative kappa would indicate agreement worse than that expected by chance.

Root Mean Square Error (RMSE)

RMSE measures the difference between the expected and the observed values from the modeling environment (Kumar and Khatri 2017). The RMSE values can be used to distinguish model performance in a training period from that of a validation period and compare the individual model performance to that of other predictive models. The RMSE of a model prediction for the estimated variable X_{pred} is defined as the square root of the mean squared error:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (X_{obs,i} - X_{pred,i})^2}{n}} \tag{9}$$

Where X_{obs} = observed values

X_{pred} = modelled values at time/place i.

n = total number of sample datasets

Receiver Operating Characteristics (ROC)

ROC is a curve that characterizes the randomly chosen probability of positive instances over negative instances (Kumar and Khatri 2017). It is a measure of different classifiers' skill with the true positives (TP) to the false-positive rates (FPR). Setting $\boldsymbol{P_{1,j}}$ as the prediction probability for the $\boldsymbol{j^{th}}$ observed event, and $\boldsymbol{P_{0,i}}$ as the prediction probability of an event for the $\boldsymbol{i^{th}}$ non-event, the ROC score, A, can be

$$A = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} I\left(P_{0,i}, P_{1,j}\right) \tag{10}$$

where $\boldsymbol{n_0}$ is the number of non-events and $\boldsymbol{n_1}$ the number of events, and the scoring rule $\boldsymbol{I}\left(\boldsymbol{P_{0,i}}, \boldsymbol{P_{1,j}}\right)$ is defined as;

$$I\left(P_{0,i}, P_{1,j}\right) = \begin{cases} 0.0 \ if \ P_{1,j} < P_{0,i} \\ 0.5 \ if \ P_{1,j} = P_{0,i} \\ 1.0 \ if \ P_{1,j} > P_{0,i} \end{cases} \tag{11}$$

In the ROC score, a hit is the selected observations are events. The proportion of all events thus selected is calculated and is known as the hit rate (HR):

$$HR = \frac{No. \ of \ TP}{No. \ of \ Event} \tag{12}$$

Some non-events may have been selected incorrectly; these are known as false positives. The proportion of non-events incorrectly chosen [the false-positive rate (FPR)] is:

$$FPR = \frac{No. \ of \ False \ positives}{No. \ of \ non \ Event} \tag{13}$$

*The ROC classifications are excellent, good, fair, poor, fail having range of [0.90–1], [0.80–0.90], [0.70–0.80], [0.60–0.70], [0.50–0.60], respectively (*Kumar and Khatri 2017).

Results

Comparison of MLP and J48 results

All three test options were used to assess MLP and J48 methods' performance, a) 10-fold cross-validation, b) percent split (66%), and the user-supplied test sets. The following section discussed the outcome of the model prediction.

Performance Evaluation of the models over Sundargarh district

The evaluations were performed for individual months and seasons to understand the prediction skill over the varying seasonal rainfall and temperature. While the prediction accuracy results exhibited that MLP and J48 were not very significantly different, but J48 had a better performance than MLP. In a similar study conducted by Gupta et al. (2011), where more attributes were analyzed and larger volumes of data were used, the prediction using J48 was better. Besides, for both the classifiers, the 10-fold cross-validation classification testing option outperforms the percentage split (66%) method for the whole district

As presented in Table 2, the prediction accuracy for both the cross-validation method (J48 and MLP) has improved during the mid-monsoon (July-August period) to late monsoon (September-October) at the same time, the prediction accuracy for the split method has declined. For all models, the dry season has comparatively lesser accuracy. If we consider kappa, the J48 cross-validation method has significantly better kappa than the other two methods; visibly after the monsoon onset, it shows a better agreement. While almost all methods have shown poor agreement during the drier summer period, results suggest the superiority of the J48 with cross-validation over the MLP percent split, MLP Cross-validation, and J48% split. The RMSE depicts that the cross-validation method has much less error compared to the percentage split method. The MLP percent split method consistently depicts the large errors across seasons. MLP has better accuracy, especially during the wet season, while less agreement to the observed condition depicts comparatively lower kappa values. Errors are

Table 2. Month-wise Performance metrics for RMSE, KAPPA, and Accuracy for the cross-validation Classifier.

Month	Accuracy		Kappa		RMSE	
	J48	MLP	J48	MLP	J48	MLP
Jan	0.68	0.61	0.56	0.44	0.69	0.70
Feb	0.63	0.63	**0.68**	0.55	**0.66**	**0.68**
Mar	0.62	0.62	0.58	0.45	**0.67**	**0.68**
Apr	0.68	0.64	0.58	0.45	**0.66**	**0.66**
May	**0.73**	0.68	0.58	0.44	**0.67**	**0.68**
Jun	0.63	**0.73**	0.59	0.45	**0.67**	**0.68**
Jul	**0.71**	**0.77**	**0.61**	0.49	0.71	0.72
Aug	**0.73**	**0.76**	**0.67**	0.50	0.72	0.73
Sep	**0.83**	**0.80**	**0.79**	**0.62**	0.70	0.72
Oct	**0.89**	**0.83**	**0.70**	**0.66**	**0.69**	0.71
Nov	**0.72**	**0.71**	**0.66**	0.51	0.71	0.72
Dec	0.60	**0.72**	0.54	0.56	0.71	0.72

Highlighted values are for Accuracy ≥ 0.70; Kappa ≥ 0.60; RMSE ≤ 0.70

more substantial for both the models during the wet period and are lower during the dry period. Since the data was analyzed monthly, J48 can be considered a more reliable predictor of malaria for the weather variables. It has comparable RMSE and higher kappa (with highest values in September = 0.79 and October =0.70) indicated that it performed better than MLP, which has Kappa value (September = 0.62 and October =0.66), respectively.

ROC score, as explained earlier, is a measure of the skill of the classifier. Evaluation with ROC requires the grouping of the prediction models into three distinct prediction categories, e.g., a) High, b) medium, and c) Low. The evaluation shows how well the three categories of events can be predicted. Table 3 lists all the four classifiers' ROC scores and provides a comparative analysis of the three-event categories, where the skill of the J48 method is comparatively better than MLP.

The J48 cross-validation method has better performance in predicting high and low events across the year. During the post-monsoon season prediction skill for high events (Sep =0.89, Oct =0.84, Nov =0.80, Dec =0.85) and (Sep =0.91, Oct =0.91, Nov =0.83, Dec =0.85) for skill for predicting 'low' events. At the same time, the percent split method has comparatively less skill.

Table 3. Month-wise ROC prediction skill scores for all four classifiers for both J48 and MLP.

Month	J48						MLP					
	Cross-validation			Split (66.0%)			Cross-validation			Split (66.0%)		
	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low
Jan	**0.79**	**0.81**	**0.82**	0.70	0.65	0.70	0.73	0.69	**0.76**	0.66	0.66	0.74
Feb	**0.84**	**0.78**	**0.83**	0.71	0.64	0.73	**0.81**	0.63	0.72	**0.77**	0.59	**0.75**
Mar	**0.77**	**0.79**	**0.78**	0.68	0.68	0.68	0.72	0.66	0.74	**0.79**	0.70	0.74
Apr	**0.77**	**0.78**	**0.77**	0.68	0.68	0.68	0.74	0.78	0.81	0.70	**0.83**	**0.79**
May	0.60	**0.76**	**0.77**	0.69	0.54	0.56	0.56	0.65	0.70	0.57	0.68	**0.86**
Jun	**0.87**	**0.77**	**0.84**	**0.79**	0.69	**0.77**	**0.79**	0.63	0.74	**0.81**	0.67	0.72
Jul	**0.82**	0.68	**0.84**	0.73	0.67	0.74	0.74	0.64	0.74	0.73	0.70	**0.75**
Aug	**0.80**	0.74	**0.83**	0.72	0.63	0.73	0.72	0.66	0.73	0.74	0.70	**0.81**
Sep	**0.89**	0.65	**0.91**	**0.88**	0.66	**0.81**	**0.80**	0.60	**0.82**	**0.88**	0.65	**0.86**
Oct	**0.84**	**0.77**	**0.91**	**0.78**	0.72	**0.83**	**0.81**	0.74	**0.86**	**0.80**	0.74	**0.85**
Nov	**0.80**	**0.80**	**0.83**	**0.77**	0.73	**0.81**	0.74	0.72	**0.80**	**0.78**	0.74	**0.82**
Dec	**0.85**	**0.82**	**0.85**	0.74	0.69	0.73	**0.75**	**0.76**	**0.79**	**0.76**	**0.75**	0.72

Highlighted values are ROC scores ≥ 0.75

MLP has even poorer results depicting skill for high events with ROC =0.56 for the May month and consistently poor throughout the year. For J48 cross-validation, the lowest value (ROC =0.65) in September and for percent split method ROC =0.54 in May, while for MLP, ROC scores for the cross-validation were ROC =0.60 in September and ROC =0.59 in February month, respectively.

Performance evaluation of the models at smaller administrative units

The block-wise analysis results suggest that the 10-fold cross-validation and the supplied test set option has yielded promising results compared to the percent split and supplied test options. Especially blocks (Kutra, Tangarapali, Sundargarh) from the central to western plain have better performance than the blocks with varying topography (Koira, Hemgiri, Bonai). The results presented in Table 4 indicate that the supplied test method has comparable performance to J48 and MLP. It has less RMSE and better accuracy, and higher kappa values. Further investigating the performance, it was observed that its accuracy of prediction is better compared to the cross-validation method, especially for central and western blocks, including Sundargarh, (Accuracy =1.0, Kappa =1.0, RMSE =0.19), Tangarpali (Accuracy =1.0, Kappa =1.0, RMSE =0.17), and Kutra (Accuracy =1.0, Kappa =1.0, RMSE =0.16). Though the supplied test options depicted smaller RMSEs and have inconsistency with accuracy and kappa (either Accuracy =1.0 or very low), it is unreliable for use in predictions.

The ROC scores suggest that the model performance is satisfactory for the central blocks like Kutra (High =0.99, Med =0.80, low =0.90), Subdega (High =0.91, Med =0.76, Low =0.86), Rajgangpur (High =0.82, Med =76, Low =84). While blocks such as Bonai (High =0.76, Med =0.63, Low =0.78), Koira (High =0.77, Med =0.75, Low =0.72), Gurundia (High =0.89, Med =0.71, Low =0.76), and Balisankara (High =0.73, Med =0.69, Low =0.74), depicts considerably lower accuracy and prediction skills. To summarize, again, J48 demonstrated a greater insight into the predictability of malaria over other methods.

Discussions

Climate sensitivity to malaria incidence

Establishing a conventional relationship between the monthly and seasonal variation of the climatic parameters to the incidents using the nominal range was beneficial to the evaluation process. The nominal range is derived from the continuous numeric data range using the R-statistical package (Mangiafico 2016). As per the derived nominal ranges (table is not included), periods of low temperature (28–30°C), low rainfall (0 –23 mm), and low relative humidity (68–82%), which is primarily the drier and cooler months of January and February are characterized by lower cases of malaria. During the months of March-April-May, a period of low rainfall (0 –23 mm), medium to a higher temperature (30°-34°C) with lower relative humidity (68–82%), there is an increase in the number of incidences of malaria, and which is noticeably due to the increase in the temperature. This also agrees with Lee et al. (2016) 's findings, a study conducted in the humid Arunachal Pradesh, India, that suggests decreasing precipitation and increasing temperature resulted in increasing malaria incidence. With the arrival of the monsoon and during the June-July-August-September, the period of high to very high rainfall (178–445 mm), high temperature (34 –38°C), and medium to high relative humidity (82–96%), the malaria incidents were further on the rise.

The malaria incidents continue to rise even after the monsoon's withdrawal and a significant drop in the temperature and humidity during the November–December period. So, there is a time lag effect of the climatic phenomenon on the incidents. This study found that extremely high temperature is a crucial trigger of the higher number of malaria incidents in Sundargarh district,

Table 4. Comparisons of Accuracy, RMSE, Kappa and ROC for all blocks for J48 cross-validation and supplied set classifiers.

Blocks	J48 Cross-Validation						J48 Supplied Test set					
	Accuracy	Kappa	RMSE	ROC			Accuracy	Kappa	RMSE	ROC		
				High	Med	Low				High	Med	Low
Hemgiri	0.57	0.40	0.89	**0.83**	0.73	0.72	**0.71**	0.16	**0.37**	**0.76**	0.74	**0.94**
Lephripara	0.69	0.33	0.88	**0.75**	**0.76**	0.72	**0.83**	0.00	**0.34**	**0.79**	0.61	0.68
Tangarpali	**0.94**	**0.55**	0.76	0.69	**0.75**	**0.75**	**1.00**	**1.00**	**0.17**	**0.81**	**0.79**	**0.85**
Sundargarh	**0.88**	0.51	0.80	0.62	0.74	0.70	**1.00**	**1.00**	**0.19**	0.45	0.69	**0.81**
Subdega	**0.71**	**0.60**	0.86	**0.91**	**0.76**	**0.86**	**0.71**	0.26	**0.36**	**0.98**	**0.84**	**0.93**
Baragaon	**0.81**	**0.63**	0.81	**0.93**	**0.77**	**0.90**	**0.79**	0.17	**0.32**	0.73	**0.76**	0.60
Balisankara	0.55	0.31	0.89	0.73	0.69	0.74	**0.79**	0.00	**0.40**	0.12	0.34	0.64
Kutra	**0.98**	0.52	0.72	**0.99**	**0.80**	**0.90**	**1.00**	1.00	**0.16**	**0.86**	**0.94**	**0.88**
Rajgangpur	0.69	0.51	0.86	**0.82**	**0.76**	**0.84**	0.63	0.00	**0.39**	**0.92**	**0.89**	**0.77**
Kuanrmunda	0.63	**0.57**	0.87	**0.97**	0.71	**0.84**	0.63	0.30	**0.41**	0.93	0.65	0.83
Nuagaon	0.58	**0.55**	0.90	**0.81**	0.67	**0.81**	0.29	0.08	0.53	0.60	0.71	0.53
Bisra	0.53	0.38	0.89	**0.88**	0.70	**0.84**	0.58	0.39	**0.49**	0.49	**0.79**	0.73
Lathikata	**0.72**	0.50	0.86	**0.88**	**0.89**	0.66	0.42	0.13	**0.50**	0.65	**0.87**	0.43
Bonai	0.57	0.50	0.91	**0.76**	0.63	**0.78**	0.46	0.18	**0.48**	0.46	**0.82**	0.59
Lahunipara	**0.70**	**0.60**	0.86	**0.90**	**0.76**	**0.94**	0.54	0.47	**0.48**	0.31	0.39	0.39
Gurundia	0.53	0.45	0.90	**0.89**	0.71	**0.76**	0.17	0.07	0.58	0.53	0.20	0.33
Koira	0.51	0.39	0.90	**0.77**	**0.75**	0.72	0.17	0.29	0.58	0.25	0.21	0.34

Highlighted values are for Accuracy ≥ 0.70; Kappa ≥ 0.55; RMSE ≤ 0.50; ROC scores ≥ 0.75

which agrees with Smith et al. (2013); and a study conducted by with Srimath-Tirumula-Peddinti et al. (2015) in Vishakhapatnam, in India, which exhibits similar climate conditions as Sundargarh. These studies found that the mosquito parasite's complete development cycle becomes faster with an increase in the temperature. Furthermore, relative humidity also affects the transmission of the malaria vector in agreement with Goswami et al. (2018), which concluded that mosquitoes survive better under high humidity conditions. During high humid seasons, the number of malaria incidents increases compared to the less humid conditions.

Model selection and model evaluation

Appropriate selection of model, algorithm, and model evaluation techniques are vital in machine learning. The evaluation intends to estimate the performance of a model or algorithm on future data. Running a learning algorithm over a training dataset with different hyperparameter settings will result in different models (Raschka 2018). Since we are interested in selecting the best-performing model, estimation helps in choosing the best model to fit the purpose though, the estimation of the absolute performance of a model is one of the most challenging tasks in machine learning (Raschka 2018). Working with small sample sizes in machine learning is acceptable but choosing the correct sampling method is vital (Cawley and Talbot 2010; Lusa 2015). Considering the sample size in this study is smaller, and for parameter optimization, 10-fold cross-validation and Leave-One-Out cross-validation are recommended as the best sampling mechanisms and generally yield better results (Cawley and Talbot 2010).

Researchers' assessments to evaluate different classifier performances (Bischl et al. 2012) recommended using the leave-one-out cross-validation method as one of the preferred prediction methods. Thus, this study put effort into assessing the 10-fold cross-validation method, which incidentally performed well. The reason should be linked to the method's data sampling and training strategy compared to the others. A list of explanations is provided below, which considers the mechanism with which the cross-validation method works.

(a) Utilized all the data samples for training and test and takes care of the multi-class issue in the percentage split method, where the sample sizes are static, and generating multiple classes means a reduction in test sets.
(b) We defined more metrics for the learning algorithm than other methods. If we have, ***n*** samples there can ***n-1*** models to predict one instance of the predictand.
(c) Through model stacking and back-propagation, models are processed in a pipeline allowing model prediction by learning from the previous model in the forward direction and feedback and model training in the backward direction. The model bias (error) is also handled better in this process.
(d) Finally, parameter fine-tuning, in which the parameters were tuned with an independent validation set, suggested the ideal number of trees in a classifier, hidden layer size (activation function) in the Neural network.

The 'model's underperformance for some blocks and months could be a factor external to climate influence. Furthermore, the interventions in place in parts of the district and other regions (Sahu et al. 2014) and the significant population's socio-economic status in the eastern belt, being tribe and access to necessary facilities is limited as also explained by Sundararajan et al. (2013). These factors influence the increase or decrease in the cases and do not truly reflect the climate's direct influence. The other reflection from the analysis is that the model's performance dips down during July and August months but improved during September and October months and declined during November and December. The variation in the model's performance is largely due to the varying topography, which affects the intra-seasonal rainfall variability and the temperature and humidity's spatial variation. For example, blocks such as Bonai, Koira, Gurundia, and Hemgiri have higher

elevation and depict considerably lower prediction accuracy. In comparison, blocks with plain land and forest cover had better performance.

A malaria early warning system and risk mapping tool is necessary to provide adequate support to the public health workers to take preparedness measures and remain prepared for any possible outbreaks in near-real-time (Connor et al. 1998; Thomson and Connor 2001; Thomson et al. 2003). Several such attempts were made, such as a spatial decision support system for Karnataka (Shekhar et al. 2017) or Kenya's operational system (Thomson et al. 2003). Statistical regression-based analysis or use of ANN or machine learning provided an opportunity to analyze the data and establish an association between the attributes. However, considering the climate dimension only in malaria, early warning is not adequate and requires a deep understanding of all other facets to establish an effective and operational warning system.

Conclusion

The climatic condition of Odisha, especially the Sundargarh district, makes it vulnerable to malaria. Monsoon rainfall, maximum surface temperature ranging from 27–40° Celsius during the summer, and relative humidity in the range of 60–85% provide a more favorable climatic condition for breeding the malaria larva in the monsoon and post-monsoon period. It was established that the increase in malaria incidents is significantly attributed to climatic factors, including temperature, humidity, and monthly rainfall variability. Among the two classifier models used, it was again concluded that J48 had shown moderately better skills over the MLP. The 10-fold Cross-validation based J48 model performance was particularly good during the early-monsoon and post-monsoon when the malaria incidents peaked. The climate is an extremely complex factor to predict, and the results provided promising signals for predicting future malaria incidents. Even though the models have shown better performance, it is constrained by datasets' non-availability for an extended period. Finer scale datasets would have provided an opportunity for deeper analysis to understand the phases and lags within a month. Furthermore, non-climatic factors such as the demography, immunity within the population, society's socio-economic structure, availability of affordable public health facilities, and other environmental modifications initiatives are strongly recommended to be factored-in while developing a malarial early warning system in a constantly changing environment.

Highlights

- The researchers analyzed the data using machine learning methods and quantify the accuracy and skill level for predicting the malaria incidences
- Among the Weka classifiers used, J48 exhibited better skill than MLP and illustrated less error and positive kappa and higher accuracy.
- 10-fold cross-validation method had better performance over the percentile split and supplied test options.
- Seasonal temperature and humidity variation had shown a better association with malaria incidents in comparison to rainfall.
- Results are encouraging for the prospect of the utilization of climate forecast for prediction of malaria incidences on a seasonal scale, which is not yet available in the region

Acknowledgments

The authors would like to acknowledge the Directorate of Public Health's contribution, Government of Odisha, for providing the malaria incident datasets. Also, the ECMWF for keeping the climate datasets free and open for public access.

Availability of data and materials

The datasets generated and analyzed during the current study are not publicly available as they are collected from government sources and due to the sensitivity but are available from the corresponding author on reasonable request.

Disclosure statement

The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding

The authors have no funding to report.

References

Al Jarullah AA 2011. Decision tree discovery for the diagnosis of type II diabetes. 2011 International conference on innovations in information technology; 2011: IEEE.

Arab A, Jackson MC, Kongoli C. 2014. Modelling the effects of weather and climate on malaria distributions in West Africa. Malar J. 13(1):126. doi:10.1186/1475-2875-13-126.

Atulbhai DK 2017. Comparision and combination of mining techniques for gene analysis to identify dengue.

[DCO] Directorate of Census Operations, Odisha. District Census Handbook, Sundargarh. SERIES-22. 2011. [Accessed 2020 17 April]. https://censusindia.gov.in/2011census/dchb/2105_PART_B_DCHB_SUNDARGARH.pdf.

Basheer IA, Hajmeer M. 2000. Artificial neural networks: fundamentals, computing, design, and application. J Microbiol Methods. 43(1):3–31. doi:10.1016/S0167-7012(00)00201-3.

Bengtsson L. 2004. Can climate trends be calculated from reanalysis data? J Geophy Res. 109(D11). doi:10.1029/2004JD004536.

Bischl B, Mersmann O, Trautmann H, Weihs C. 2012. Resampling methods for meta-model validation with recommendations for evolutionary computation. Evol Comput. 20(2):249–275. doi:10.1162/EVCO_a_00069.

Bomblies A. 2012. Modeling the role of rainfall patterns in seasonal malaria transmission. Clim Change. 112 (3–4):673–685.

Bui Q-T, Nguyen Q-H, Pham VM, Pham MH, Tran AT. 2019. Understanding spatial variations of malaria in Vietnam using remotely sensed data integrated into GIS and machine learning classifiers. Geocarto Int. 34 (12):1300–1314.

Cawley GC, Talbot NL. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. J Machine Learning Res. 11:2079–2107.

Chatterjee C, Sarkar RR. 2009. Multi-step polynomial regression method to model and forecast malaria incidence. PLoS One. 4(3):e4726.

Connor SJ, Thomson MC, Flasse SP, Perryman AH. 1998. Environmental information systems in malaria risk mapping and epidemic forecasting. Disasters. 22(1):39–56.

Dangare CS, Apte SS. 2012. Improved study of heart disease prediction system using data mining classification techniques. Int J Comp Appl. 47(10):44–48.

Devi NP, Jauhari RK. 2006. Climatic variables and malaria incidence in Dehradun, Uttaranchal, India. J Vector Borne Dis. 43(1):21.

Dhiman RC, Pahwa S, Dash AP. 2008. Climate change and malaria in India: interplay between temperature and mosquitoes. Regional Health Forum. 12(1):27–31.

Dhiman RC, Pahwa S, Dhillon GPS, Dash AP. 2010. Climate change and threat of vector-borne diseases in India: are we prepared? Parasitol Res. 106(4):763–773.

Gallup JL, Sachs JD. 2001. The economic burden of malaria. Am. J. Trop. Med. Hyg. 64(1_suppl):85–96.

Githeko AK, Lindsay SW, Confalonieri UE, Patz JA. 2000. Climate change and vector-borne diseases: a regional analysis. Bull World Health Organ. 78:1136–1147.

Goswami S, Saxena A, Singh KJ, Chandra S, Cleal CJ. 2018. An appraisal of the Permian palaeobiodiversity and geology of the Ib-River Basin, eastern coastal area, India. Journal of Asian Earth Sciences. 157:283–301.

Guo C, Yang L, Ou C-Q, Li L, Zhuang Y, Yang J, Zhou Y-X, Qian J, Chen P-Y, Liu Q-Y. 2015. Malaria incidence from 2005–2013 and its associations with meteorological factors in Guangdong, China. Malar J. 14(1):116.

Guo P, Liu T, Zhang Q, Wang L, Xiao J, Zhang Q, Luo G, Li Z, He J, Zhang Y. 2017. Developing a dengue forecast model using machine learning: a case study in China. PLoS Negl Trop Dis. 11(10):e0005973.

Gupta R. 1996. Correlation of rainfall with upsurge of malaria in Rajasthan. J Assoc Physicians India. 44(6):385–389.

Gupta S, Kumar D, Sharma A. 2011. Performance analysis of various data mining classification techniques on healthcare data. Int J Comp Sci Inf Technol. 3(4):155–169.

Hay SI, Guerra CA, Tatem AJ, Atkinson PM, Snow RW. 2005. Urbanization, malaria transmission and disease burden in Africa. Nat. Rev. Microbiol. 3(1):81–90.

Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JC, Wides R. 2002. The genome sequence of the malaria mosquito Anopheles gambiae. science. 298 (5591):129–149.
Hornick M. 2009. Data mining agents for efficient hardware utilization. Google Patents.
Imai C, Armstrong B, Chalabi Z, Mangtani P, Hashizume M. 2015. Time series regression model for infectious disease and weather. Environ Res. 142:319–327.
Jolivet R, Grandin R, Lasserre C, Doin MP, Peltzer G. 2011. Systematic InSAR tropospheric phase delay corrections from global meteorological reanalysis data. Geophys Res Lett. 38:17.
Kakmeni FMM, Guimapi RY, Ndjomatchoua FT, Pedro SA, Mutunga J, Tonnang HE. 2018. Spatial panorama of malaria prevalence in Africa under climate change and interventions scenarios. Int J Health Geogr. 17(1):1–13.
[NVBDCP]. National Vector Borne Disease Control Programme. n.d.. Malaria Situation Report. In: Services DGoH, editor. Ministry of Health & Family Welfare, Govt. of India: Malaria situation in India. 2020 28. https://nvbdcp.gov.in/index4.php?lang=1&level=0&linkid=564&lid=3867.10
Kannan KP. 2017. Interrogating inclusive growth: poverty and inequality in India. Routledge.
Kaur G, Chhabra A. 2014. Improved J48 classification algorithm for the prediction of diabetes. Int J Comp Appl. 98:22.
Kim YM, Park JW, Cheong HK. 2012. Estimated effect of climatic variables on the transmission of Plasmodium vivax malaria in the Republic of Korea. Environ Health Perspect. 120(9):1314–1319.
Korting TS. 2006. C4. 5 algorithm and multivariate decision trees. Brazil: Image Processing Division, National Institute for Space Research–INPE Sao Jose dos Campos–SP.
Kovats RS, Bouma MJ, Hajat S, Worrall E, Haines A. 2003. El Niño and health. The Lancet. 362(9394):1481–1489.
Kumar N, Khatri S 2017. Implementing WEKA for medical data classification and early disease prediction. 2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT); 2017: IEEE.
Landis JR, Koch GG. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics. 363–374.
Leal Filho W, Bönecke J, Spielmann H, Azeiteiro UM, Alves F, De Carvalho ML, Nagy GJ. 2018. Climate change and health: an analysis of causal relations on the spread of vector-borne diseases in Brazil. J Clean Prod. 177:589–596.
Lee E, Burkhart J, Olson S, Billings AA, Patz JA, Harner EJ. 2016. Relationships of climate and irrigation factors with malaria parasite incidences in two climatically dissimilar regions in India. J Arid Environ. 124:214–224.
Leo J, Luhanga E, Michael K. 2019. Machine learning model for imbalanced cholera dataset in Tanzania. Sci World J. 2019.
Lingala MA. 2017. Effect of meteorological variables on Plasmodium vivax and Plasmodium falciparum malaria in outbreak prone districts of Rajasthan, India. J Infect Public Health. 10(6):875–880.
Lowassa A, Mazigo HD, Mahande AM, Mwang'onde BJ, Msangi S, Mahande MJ, Kimaro EE, Elisante E, Kweka EJ. 2012. Social economic factors and malaria transmission in Lower Moshi, northern Tanzania. Parasites vectors. 5 (1):1–9.
Lusa L. 2015. Joint use of over-and under-sampling techniques and cross-validation for the development and assessment of prediction models. BMC Bioinformatics. 16:(1):1–10.
Macherera M, Chimbari MJ. 2016. A review of studies on community based early warning systems. Jàmbá: J of disaster risk stud. 8(1).
Mahakur PK, Nayak NC. 2019. Intrastate income inequalities in Odisha: examining decomposition by regions and broad sectors. Odisha Econ Dis Ser. 1.
Mahmud SH, Hossin MA, Ahmed MR, Noori SR, Sarkar MN 2018. Machine learning based unified framework for diabetes prediction. Proceedings of the 2018 International Conference on Big Data Engineering and Technology; 2018.
Mangiafico S. 2016. Summary and analysis of extension program evaluation in R, version 1.15. 0. New Brunswick (NJ): Rutgers Cooperative Extension. https://rcompanionorg/handbook/
Marcos R, González-Reviriego N, Torralba V, Soret A, Doblas-Reyes FJ. 2019. Characterization of the near surface wind speed distribution at global scale: ERA-Interim reanalysis and ECMWF seasonal forecasting system 4. Climate Dynamics. 52(5–6):3307–3319.
McHugh ML. 2012. Interrater reliability: the kappa statistic. Biochemia Medica. 22(3):276–282.
Mello-Román JD, Mello-Román JC, Gomez-Guerrero S, García-Torres M. 2019. Predictive models for the medical diagnosis of dengue: a case study in Paraguay. Comput Math Methods Med. 2019.
Mishra G. 2003. Hospital based study of malaria in Ratnagiri district, Maharashtra. J Vector Borne Dis. 40(3/4):109.
Mouchet J, Manguin S, Sircoulon J, Laventure S, Faye O, Onapa A, Carnevale P, Julvez J, Fontenille D. 1998. Evolution of malaria in Africa for the past 40 years: impact of climatic and human factors. J. Am. Mosq. Control Assoc. 14(2):121–130.
Neter J, Kutner MH, Nachtsheim C, Wasserman W. 1996. Applied linear statistical models. WCB McGraw-Hill.
Ngarakana-Gwasira ET, Bhunu CP, Masocha M, Mashonjowa E. 2016. Assessing the role of climate change in malaria transmission in Africa. Malar Res Treat, 2016.

Nicholson C n.d. A Beginner's Guide to Multilayer Perceptrons (MLP). [Accessed Apr 2020 20]. https://pathmind.com/wiki/multilayer-perceptron#three.%20https://pathmind.com/wiki/multilayer-perceptron#three
Olayinka T, Chiemeke S. 2019. Predicting paediatric malaria occurrence using classification algorithm in data mining. J Advan Mathematics Comp. Science.1–10.
Parham PE, Michael E. 2010. Modeling the effects of weather and climate change on malaria transmission. Environ Health Perspect. 118(5):620–626.
Parker WS. 2016. Reanalyses and observations: 'What's the difference? Bull Am Meteorol Soc. 97(9):1565–1572.
Pramanik M, Udmale P, Bisht P, Chowdhury K, Szabo S, Pal I. 2020. Climatic factors influence the spread of COVID-19 in Russia. Int J Environ Health Res. doi:10.1080/09603123.2020.1793921
Quinlan J. 2014. C4. 5: programs for machine learning. Elsevier.
Raschka S. 2018. Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv. p. 181112808.
Rejeki DSS, Nurhayati N, Budi A, Murhandarwati EEH, Kusnanto H. 2018. A time series analysis: weather factors, human migration and malaria cases in endemic area of Purworejo, Indonesia, 2005–2014. Iran J Public Health. 47 (4):499.
Sabarinathan V, Sugumaran V. 2014. Diagnosis of heart disease using decision tree. Int J Res Comp Appl Inf Technol. 2(6):74–79.
Sahu S, Gunasekaran K, Raju H, Vanamail P, Pradhan M, Jambulingam P. 2014. Response of malaria vectors to conventional insecticides in the southern districts of Odisha State, India. Indian J Med Res. 139(2):294.
Segun OE, Shohaimi S, Nallapan M, Lamidi-Sarumoh AA, Salari N. 2020. Statistical modelling of the effects of weather factors on malaria occurrence in Abuja, Nigeria. Int J Environ Res Public Health. 17(10):3474.
Shakil KA, Anis S, Alam M 2015. Dengue disease prediction using weka data mining tool. arXiv preprint arXiv: 150205167.
Sharma V, Kumar A, Lakshmi Panat D, Karajkhede G. 2015. Malaria outbreak prediction model using machine learning. Int J Advan Res Comp Eng Technol. 4:12.
Shekhar S, Yoo E, Ahmed S, Haining R, Kadannolly S. 2017. Analysing malaria incidence at the small area level for developing a spatial decision support system: a case study in Kalaburagi, Karnataka, India. Spat Spatiotemporal Epidemiol. 20:9–25.
Shimaponda-Mataa NM, Tembo-Mwase E, Gebreslasie M, Achia TN, Mukaratirwa S. 2017. Modelling the influence of temperature and rainfall on malaria incidence in four endemic provinces of Zambia using semiparametric Poisson regression. Acta Trop. 166:81–91.
Smith DL, Perkins TA, Tusting LS, Scott TW, Lindsay SW. 2013. Mosquito population regulation and larval source management in heterogeneous environments. PloS One. 8(8):e71247.
Srimath-Tirumula-Peddinti RCPK, Neelapu NRR, Sidagam N. 2015. Association of climatic variability, vector population and malarial disease in district of Visakhapatnam, India: a modeling and prediction analysis. PLoS One. 10(6):e0128377.
Sundararajan R, Kalkonde Y, Gokhale C, Greenough PG, Bang A. 2013. Barriers to malaria control among marginalized tribal communities: a qualitative study. PloS One. 8(12):e81966.
Tatem AJ, Hay SI. 2004. Measuring urbanization pattern and extent for malaria research: a review of remote sensing approaches. J. Urban Health. 81(3):363–376.
Thomson M, Indeje M, Connor S, Dilley M, Ward N. 2003. Malaria early warning in Kenya and seasonal climate forecasts. The Lancet. 362(9383):580.
Thomson MC, Connor SJ. 2001. The development of malaria early warning systems for Africa. Trends Parasitol. 17 (9):438–445.
Van Lieshout M, Kovats R, Livermore M, Martens P. 2004. Climate change and malaria: analysis of the SRES climate and socio-economic scenarios. Global Environ Change. 14(1):87–99.
Viera AJ, Garrett JM. 2005. Understanding interobserver agreement: the kappa statistic. Fam Med. 37(5):360–363.
Weka. A machine learning workbench. Proceedings of ANZIIS'94-Australian New Zealnd Intelligent Information Systems Conference; 1994: IEEE.
Williams HA, Bloland PB, Council NR, Population Co. 2003. Malaria control during mass population movements and natural disasters. National Academies Press.
Witten IH, Frank E. 2002. Data mining: practical machine learning tools and techniques with Java implementations. Acm Sigmod Record. 31(1):76–77.
World Health Organization [WHO]. 2018. World malaria report 2018. Geneva:World Health Organization.
World Health Organization [WHO]. 2019. World malaria report 2019. Geneva:World Health Organization.
Yadav K, Dhiman S, Rabha B, Saikia P, Veer V. 2014. Socio-economic determinants for malaria transmission risk in an endemic primary health centre in Assam, India. Infect. Dis. poverty. 3(1):1–8.
Yao J, Tan CL, Poh H-L. 1999. Neural networks for technical analysis: a study on KLCI. Int J Theoretical Appl Finance. 2(02):221–241.
Zia UA, Khan N. 2017. Predicting diabetes in medical datasets using machine learning techniques. Int J Sci Eng Res. 8:5.