Comparative Analysis of Machine Learning Classifiers for the Prediction of Malaria Incidence Attributed to Climatic Factors

Pallavi Mohapatra (mohapatra.pallavi@gmail.com)
Asian Institute of Technology https://orcid.org/0000-0003-3491-7469
N.K. Tripathi
Asian Institute of Technology
Indrajit Pal
Asian Institute of Technology
Sangam Shrestha
Asian Institute of Technology

Research

Keywords: Machine Learning, Malaria prediction, J48 Decision Tree, WEKA, Multilayer Perceptron

Posted Date: August 7th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-52162/v1

Research Article

Comparative Analysis of Machine Learning Classifiers for the Prediction of Malaria Incidence Attributed to Climatic Factors

Pallavi Mohapatra*

Remote Sensing and Geographic Information System, Asian Institute of Technology, Pathum Thani, Thailand,

Email: mohapatra.pallavi@gmail.com

N. K. Tripathi

Remote Sensing and Geographic Information System, Asian Institute of Technology, Pathum Thani, Thailand,

Email: nitinkt@ait.ac.th

Indrajit Pal

Disaster Preparedness Mitigation and Management, Asian Institute of Technology, Pathum Thani, Thailand.

Email: indrajit-pal@ait.ac.th

Sangam Shrestha

Water Engineering and Management, Asian Institute of Technology, Pathum Thani, Thailand;

Email: sangam@ait.ac.th

* Author for correspondence (e-mail: mohapatra.pallavi@gmail.com; Tel: +66-873369790)

Abstract

Background: India has a rising rate of malaria as well as a high mortality rate despite awareness and efforts being focused on the issue. Some regions are profoundly affected than others, such as in Odisha, where the prevalence of malaria is nearly a third of the whole country. This study investigated the influence of climate factors on the incidence of malaria in the Sundargarh district in the state of Odisha, India.

Methods: Block-wise observed station rainfall data was sourced from the Special Relief Commissioners' (SRC) web portal. Gridded surface maximum temperature and relative humidity data were accessed from the European Center for Medium-range Weather Forecast (ECMWF) reanalysis data archive. Malaria incident data were collected from the Directorate of Public Health, Government of Odisha. WEKA machine learning tool with two classifier techniques, Multi-Layer Perceptron (MLP) and J48 with 10-fold cross-validation, percentile split (66%), and supplied test options, were used for the Malaria prediction. A comparative analysis was carried out on both techniques to ascertain the superior model amongst the two, concerning the prediction accuracy of malaria in the context of a varying climate. Classifier accuracy, Root Mean Square Error (RMSE), Kappa, and ROC scores were the indicators used for the analysis.

Results: The results suggested that J48 had exhibited a better skill to MLP and illustrated less error with a positive kappa. In particular, the 10-fold cross-validation method had better performance over the percentile Spilt (66%) 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. Further, seasonal temperature and humidity variation had shown a better association with malaria incidents in comparison to rainfall.

Conclusion: The performance of the machine learning methods for Sundargarh was particularly better during the monsoon and post-monsoon when the events are at the peak. The results were encouraging for the utilization of climate forecast for prediction of malaria incidences. It is thus recommended that the J48 classifier machine learning technique could be adopted for the development of malaria early warning system.

Keywords: Machine Learning, Malaria prediction, J48 Decision Tree, WEKA, Multilayer Perceptron

---

Background

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 [1]. According to Kovats et al. [2], malaria is characterized by seasonal transmission and distribution of vectors and is influenced by seasonal climatic variations [3]. This is because both vectors and parasites tend to be sensitive to changes in atmospheric temperature and moisture [4]. 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 [5]. The examination of how climate conditions could affect the spreading of malaria can be approached by closely monitoring various aspects that change in the climate and the surrounding environment. Van et al. [6] examined the spatio-temporal effects of climate change on malaria. They established that significant changes in the temperature and rainfall patterns could lead to an increase in the spreading of malaria.

Malaria prevalence depends on the parasite *Plasmodium* and population dynamics of *Anopheles* mosquito [7]. The development, as well as the survival rates of both *Plasmodium* parasites and the *Anopheles* mosquitoes, is dependent on weather. As Kakmeni et al. [8] explain, more specifically, the temperature is key to the persistence of these parasites. Current evidence suggests that inter-decadal and inter-annual variability of the climate have a direct effect on the epidemiology of some of the critical vector-borne diseases [2]. Odisha, an eastern coastal state in India, has the maximum number of incidents of malaria and causalities since 2014, compared to other states (provinces) of India, as per the statistics (as shown in Figure 1) provided by the National Vector Borne Disease Control Programme [9]. The figure shows only the states with the highest number of incidents. The geographical positioning of the state made it susceptible to climate extremes and adversely affecting human health.

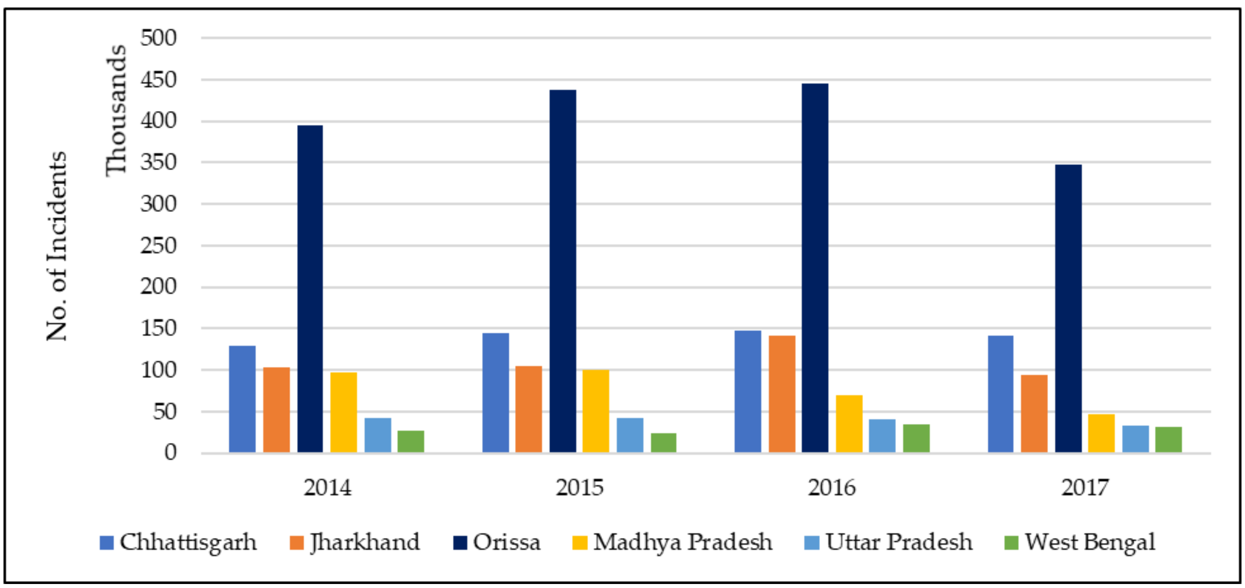


Figure 1. Malaria incidents across six different states of India for the period of 2014-2017.

(Source: NVBDCP Malaria situation reports)

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 possibility 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. Block-wise cumulative incidence map for the district is presented in Figure 2, which suggest the eastern region is severely affected by Malaria, in particular. The study has evaluated the efficiency of the machine learning algorithms by determining the accuracy level at which they were able to predict the malaria incidents. Further, the findings would also encourage better utilization of climate forecasts to predict potential malaria risk.

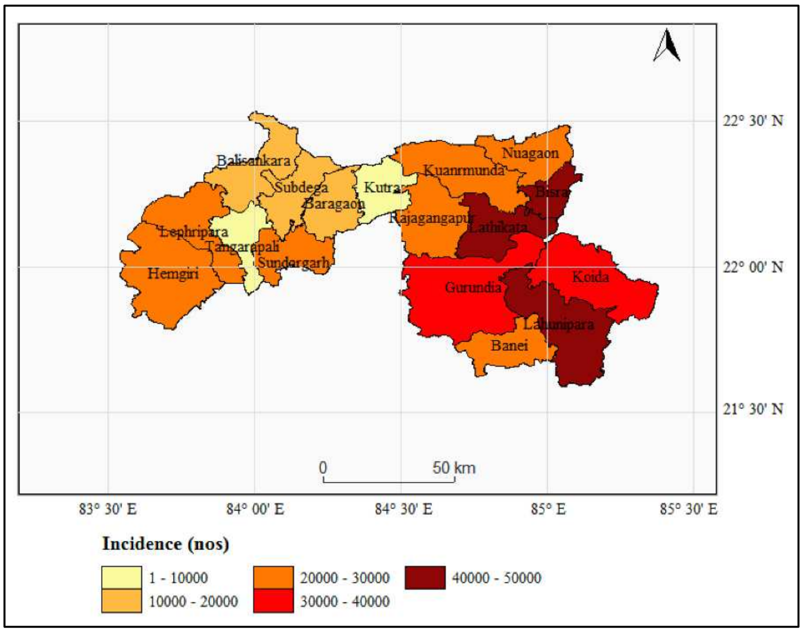


Figure 2. Cumulative malaria incidence map for the period 2002 to 2017 (Data Source: Directorate of Public Health, Odisha)

Malaria as a Public Health Concern

There were 219 million cases of malaria globally in 2017 [1], and an estimated 228 million cases of malaria occurred worldwide in 2018 [10]. The burden was most substantial in the African region, where an estimated 93% of all malaria deaths occurred, and in children aged under five years, who accounted for 61% of all deaths [1]. 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 [10]. Malaria is prevalent in eastern, central, and north-eastern states, especially in ethnic groups, usually, dominate these tribal areas. Inequality and poverty in this area play a crucial role in the spreading as well [11,12]. The most vulnerable community to malaria is the tribal populations in India habitually reside in remote areas with complex topography and dense forest with limited to no access to basic facilities [13].

Influence of Climate on Malaria

Statistical methods were used by several researchers to investigate the association of climatic factors and malaria incidents which included the multiple polynomial regression to model malaria incidents in India [14], semi-parametric Poisson distribution methods to model the influence of temperature and rainfall on malaria incidence in Zambia [15], distributed non-linear lag model to associate malaria to meteorological factors in China [16], hierarchical Bayesian framework to model effects of weather and climate on malaria distributions in West Africa [17] and the time series regression models [18,19]. All the models have shown reasonable skills over the respective regions. Neter et al. [20] recommend the use of Multiple Linear Regression (MLR) in the analysis of 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 [21] 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. [22] demonstrated that using neural network for data analysis could detect all possible interactions among the predictors.

This research acknowledges the existing problem of the malaria epidemic in the region and the influence of increased frequency of extreme climate events such as floods, heatwaves, drought, which contribute further to the escalation of malaria spreading. Malaria is no doubt a significant threat to human life, and climate variability plays a key role in the survival and abundance of the disease vectors. The researchers analyzed the data using the machine learning methods and quantify the accuracy and skill level for predicting the incidence.

Methods

Study Area

The statistics from the National Vector Borne Disease Control Programme revealed that Odisha, a coastal province in India, records the maximum number of causalities due to vector-borne diseases and especially from malaria [9]. This research targeted the Sundargarh district, which records one of the highest numbers of cases of malaria incidents in the state. The geographical location of Odisha made it susceptible to increased occurrences of climate extremes and adversely affecting human 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 area 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 [23].

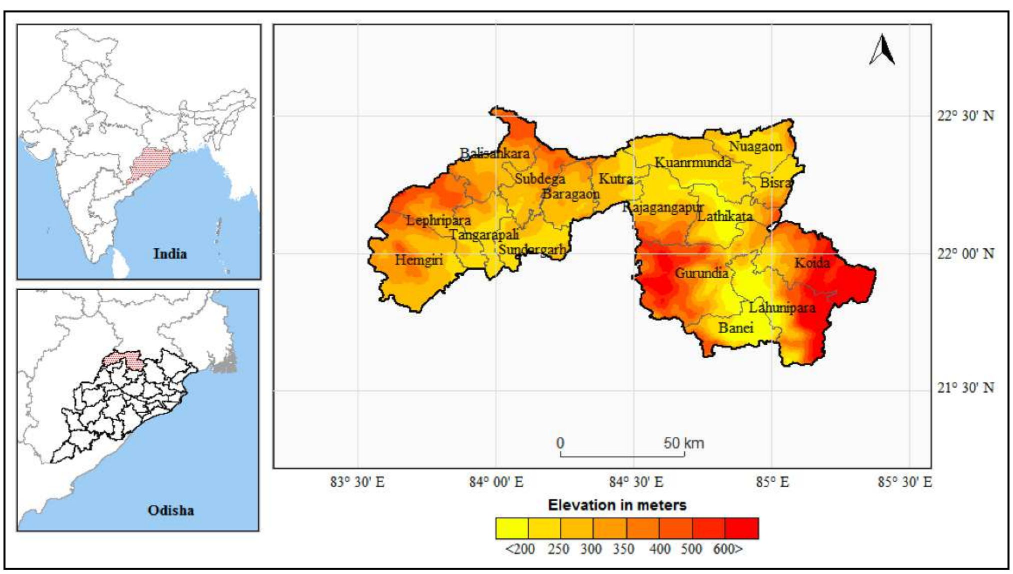


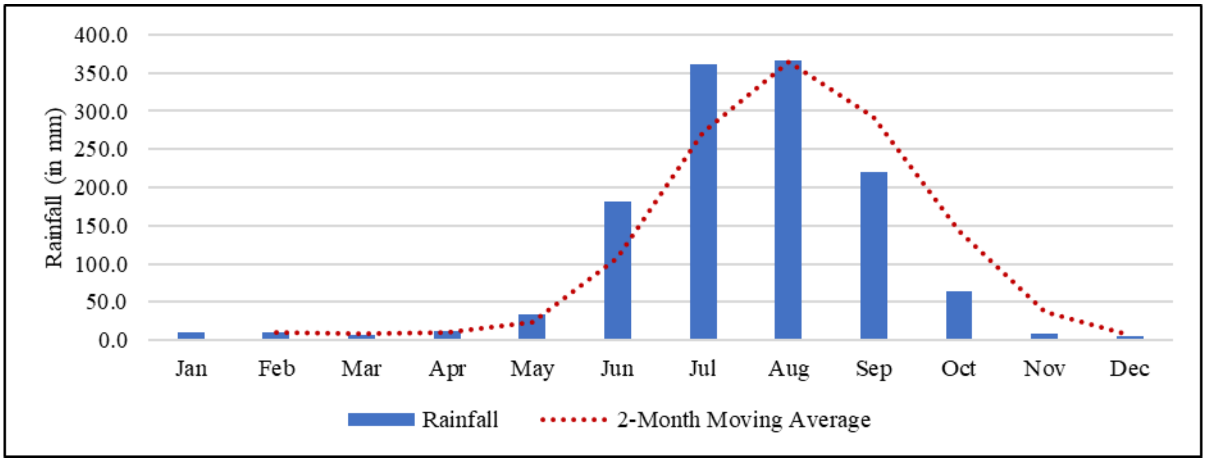
Figure 3. Study area with elevation for Sundargarh district in Odisha, India

Topography

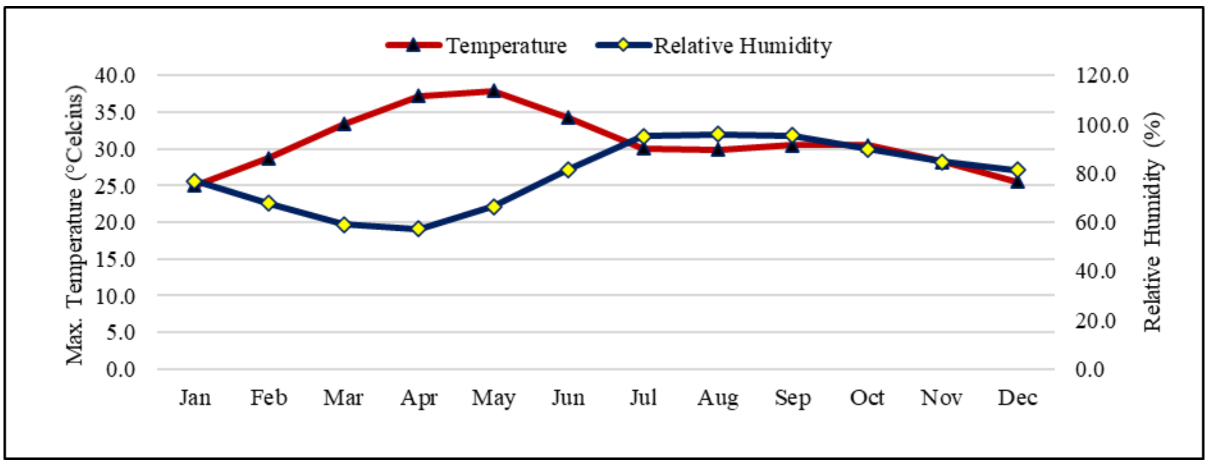
The district exhibits an ideal ecological condition for the malaria transmission topographically with its undulating uplands intersected by forested hills and widely diversified tracts of mountains. The areas covered by western blocks are long undulating tracts of about 700 ft. (213 mt.) above the sea level, dotted with hill-ranges and isolated peaks of considerable height. At the same time, 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

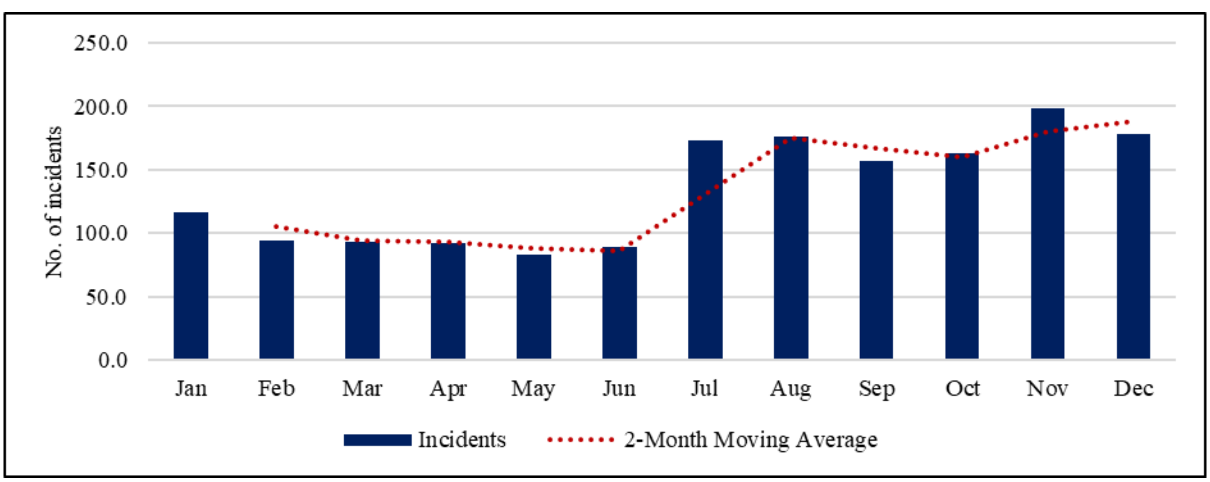
From the southwest monsoon, the area received rainfall between June and September and characterized as a tropical humid climate region (shown in Figure 4 (a) and (b)). The average annual temperature ranges between 22°C and 27°C and the average annual rainfall ranges between 1600 and 2000 mm. The weather 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.


400.0
350.0
300.0
250.0
200.0
150.0
100.0
50.0
0.0
Rainfall (in mm)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Rainfall
2-Month Moving Average

(a)


Temperature
Relative Humidity
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
Max. Temperature (°Celcius)
120.0
100.0
80.0
60.0
40.0
20.0
0.0
Relative Humidity (%)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

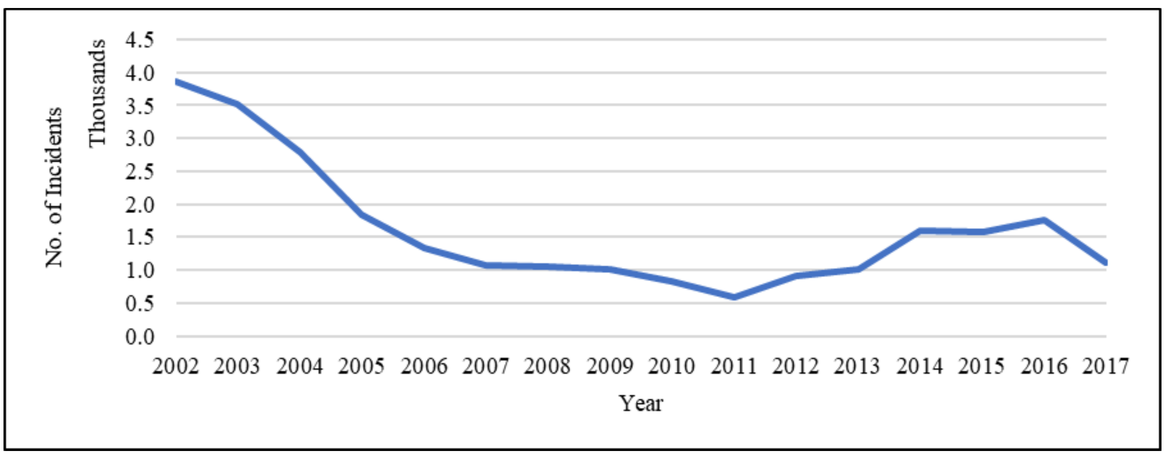
(b)


250.0
200.0
150.0
100.0
50.0
0.0
No. of incidents
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Incidents
2-Month Moving Average

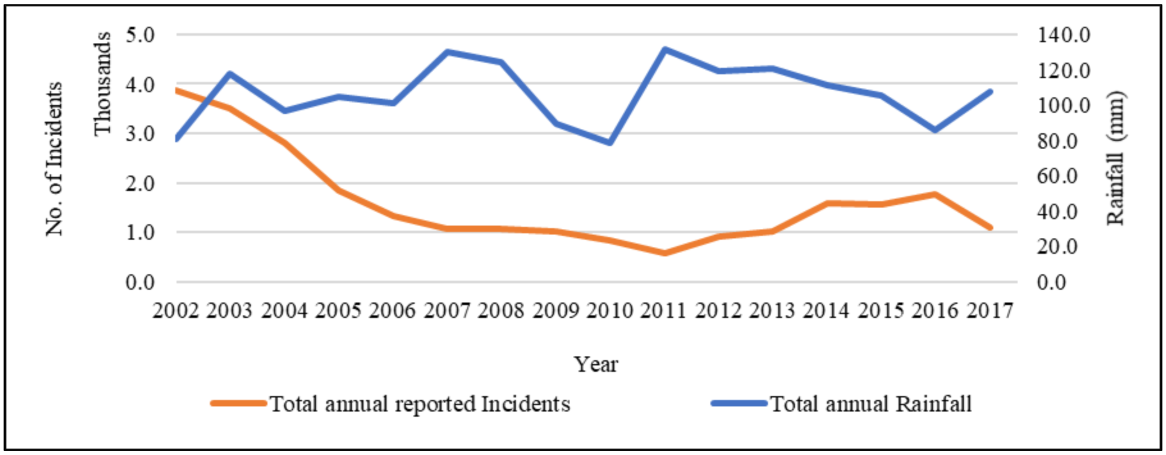
(c)

Figure 4. Average monthly rainfall (in mm) (a), maximum surface temperature (in °C), relative humidity (in %) (b), and incidents reported (c) for the period of 2002 to 2017.

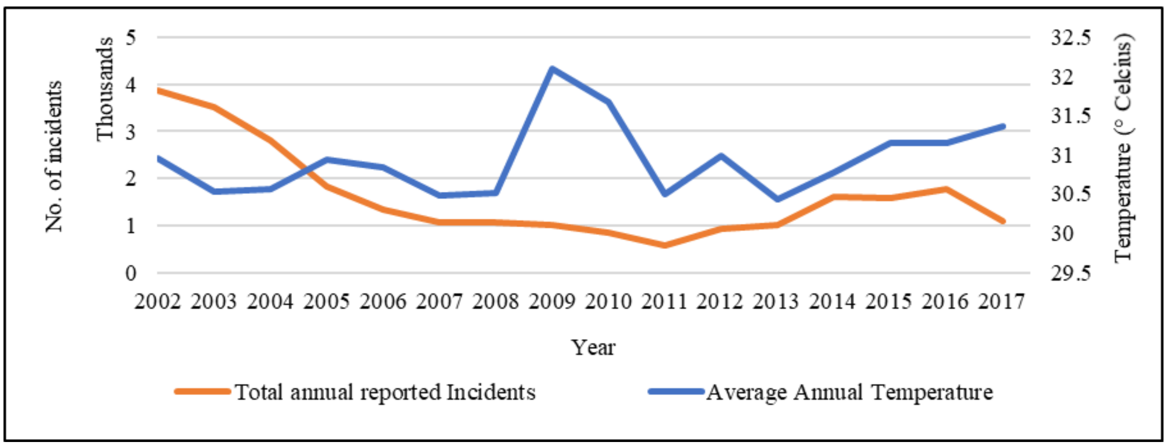
Seasonal variation in temperature plays an important role as far as vector diseases, and their transmission is concerned. According to Servadio et al. [24], the most significant effect of climate change on the transmission of malaria would be felt at the extreme temperature ranges. This reiterates the importance of studying climate variability and determining how its changes can affect the transmission of malaria, not just in places that are already affected by the disease, but also in fresher areas. Figure 4 provides the seasonal influence of the incidents to the rainfall, maximum temperature, and relative humidity, and Figure 5 shows the trend of incidents for 2002 to 2017 and its relationship with the three climate factors under consideration in this study.



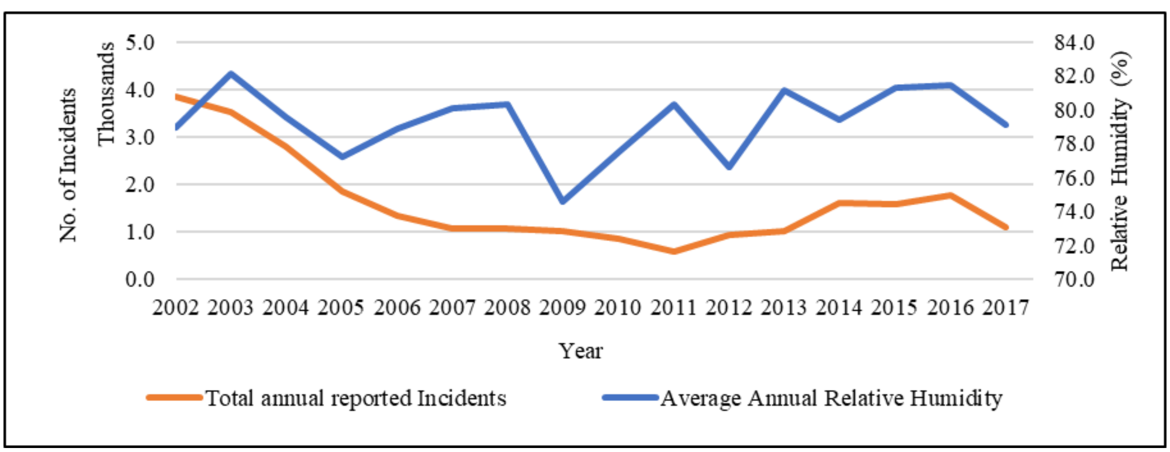
(a)


Thousands
No. of Incidents
Rainfall (mm)
5.0
4.0
3.0
2.0
1.0
0.0
140.0
120.0
100.0
80.0
60.0
40.0
20.0
0.0
2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
Year
Total annual reported Incidents
Total annual Rainfall

(b)


Thousands
No. of incidents
Temperature (° Celcius)
5
4
3
2
1
0
32.5
32
31.5
31
30.5
30
29.5
2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
Year
Total annual reported Incidents
Average Annual Temperature

(c)


Thousands
No. of Incidents
Relative Humidity (%)
5.0
4.0
3.0
2.0
1.0
0.0
84.0
82.0
80.0
78.0
76.0
74.0
72.0
70.0
2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
Year
Total annual reported Incidents
Average Annual Relative Humidity

(d)

Figure 5. The trend of annual incidents (a); and comparison with average annual rainfall (b); temperature (c); and relative humidity (d) for the period of 2002 to 2017

Data Used

The variables used in this research were climate parameters like 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 climate parameters. Many studies have demonstrated use of gridded reanalysis data in the absence of actual ground observation [25, 26, 27, 28] 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 2002-2017. Table 1 summarizes the data used in the study.

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

Data Preparation

The data received 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 spread of the variables. The next step in the analysis was changing the historical data into range values, and turning it from numerical to nominal (Low, Medium, High, and Very High). A sample conversion is shown in Table 2.

Table 2. Conversion of Numeric data to Nominal data

Year	RF (mm)	$T2_{max}$ (°C)	RH (%)	Incidents	RF.	$T2_{max}$	R.H.	Incidents
		Numeric Data				Nominal Data		
2002	24	24.7	82.6	111	L	L	L	L

2002	15	32.8	66.1	129	L	M	L	M
2002	43	36.3	72.0	175	M	H	L	M
2002	308	28.8	98.1	250	H	L	VH	H
2002	169.5	29.5	96.9	195	H	L	H	H
2002	84.4	30.5	92.9	177	M	M	M	M

L=Low, M=medium, H=High, VH=Very High

Weka Machine Learning Tool

Waikato Environment for Knowledge Analysis (WEKA) is a collection of machine learning algorithms that accurately perform data mining tasks [29]. WEKA contains tools that facilitate data preparation, regression, classification, association rules mining, clustering, and visualization. WEKA, through its machine learning platform, enables the algorithm to learn about data as samples and with or without the interference of any other explicit programs [30,31]. More detail about the tool is available at https://www.cs.waikato.ac.nz/~ml/weka/. Multilayer Perceptron (MPL) and J48 classifier techniques in the Weka tool recently being used in successfully predicting malaria incidents [32-34]. Researchers around the globe also used it for prediction of dengue [35-38] and other public health issues such as Cholera [39], diabetes [40-42], heart diseases [43, 44].

Multiple Layer Perceptron

A Multilayer Perceptron (MLP) is a class of feed-forward artificial neural networks [45]. 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. ANN has, for a long time, been a robust perceptive classifier for tasks not just in medical diagnosis, but also for early detection of diseases [46]. MLP uses a supervised learning technique that is referred to as propagation for training the network [45]; it is a modification of 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 [46]. Which can be represented in the following form:

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

Where $\boldsymbol{w}$ denotes the vector of weights, $\boldsymbol{x}$ is the vector of inputs, $\boldsymbol{b}$ is the bias, and $\boldsymbol{\varphi}$ 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 [46]. 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 [45, 47]. Dangare & Apte [44] 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 be used for the prediction of 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 [45]. In other words, the classification of data would be as perfect as possible. The objective of the J48 classification is to reduce the impurity or uncertainty in data as much as possible. A subset of data is 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., a measure of the disorder of the data [45, 47, 48]. The Entropy of $\vec{\boldsymbol{E}}$ is calculated as:

$$Entropy(\vec{E}) = -\sum_{j=1}^{n} \frac{|E_j|}{|\vec{E}|} \log \frac{|E_j|}{|\vec{E}|} \quad (2)$$

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

$$Entropy(j \mid \vec{E}) = \frac{|E_j|}{|\vec{E}|} \log \frac{|E_j|}{|\vec{E}|} \quad (3)$$

and finally, the gain is

$$gain(\vec{E}, j) = entropy\ (\vec{E} - Entropy\ (j|\vec{E})) \quad (4)$$

The aim is to maximize the gain, dividing by overall Entropy due to split argument $\vec{\boldsymbol{E}}$ by value j.

Predictive Modeling using MLP and J48

The steps followed for the predictive modelling using MLP and J48 are presented in Figure 6. The climate datasets collected from the respective sources were reprocessed to monthly scale as the prediction of malaria incidents was expected to be carried out at a monthly time scale. The numerical monthly malaria incident and climate data were then, transformed to nominal range, before they were fed to the MLP and J48 classifier models. The WEKA tool [30] was used as a base platform for all the analyses. The datasets were split into two sets; first set for training of the model and the second set for testing or the prediction. Different test options used include; (a)

10-fold cross-validation method, in which all samples were divided as ten equal sets, from which 1 set is used for testing, and the rest nine sets for the training of the model; (b) percentage split method, in which the data is distributed as a percent of the total number of samples with 34% data are used for testing and rest 66% data are for training; and (c) a supplied test set, which enables users to decide on the distribution of the samples for training and prediction. Various indicators including the RMSE, Kappa, ROC, accuracy was used to evaluate the performance of the techniques used for the prediction. From the investigation, the better performing technique with the most appropriate test option was identified. Further, the technique and test methods would be used as a malaria prediction engine for the prediction of malaria though a Malaria early warning system.

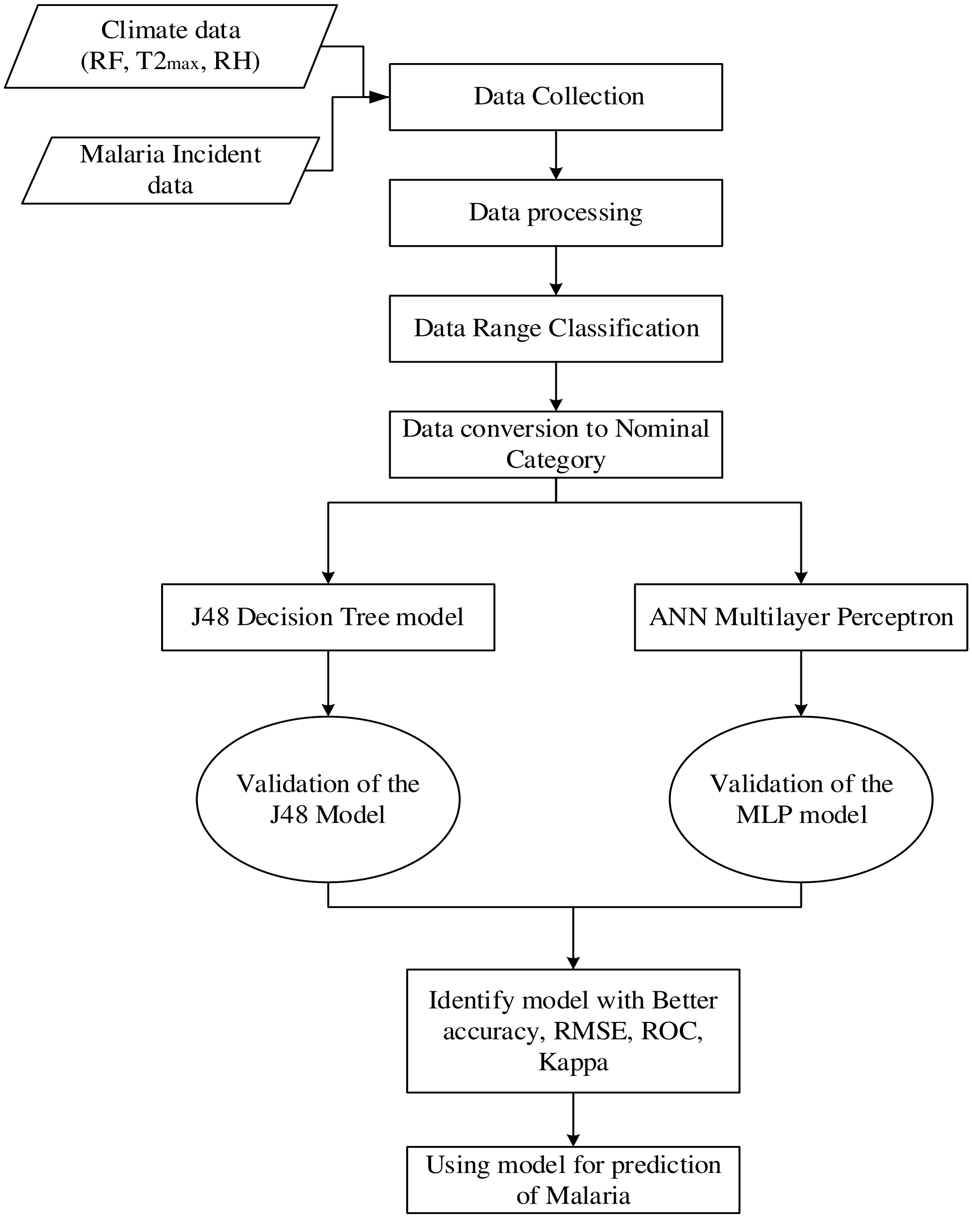


Figure 6. Detailed Methodology for Predictive Modeling of Malaria

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 performance of the machine learning analysis methods was evaluated though different indicators which 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 its significance. The confusion matrix (Table 3 and Table 4) provides a simplified structure of the representation of the observed and exptected samples to segregate the classifications into four classes: True Positives ($\boldsymbol{a}$), False Positives ($\boldsymbol{b}$), False Negatives ($\boldsymbol{c}$) and True Negatives ($\boldsymbol{d}$).

Table 3. Confusion Matrix for observed agreement

		Observed		
Expected		Positive	Negative	Total
	Positive	***a*** (TP)	***b*** (FP)	***a+b***
	Negative	***c*** (FN)	***d*** (TN)	***c+d***
Total		***a+c***	***b+d***	***N***

TP=True positives; FP = False Positives; FN=False Negatives; TN=True Negatives

The observed agreement is the frequency with which the two variants (observed and expected) agreed. From the confusion matrix, the observed agreement can be determined as:

$$Observed\ agreement = \frac{\boldsymbol{a} + \boldsymbol{d}}{\boldsymbol{N}} \tag{5}$$

Table 4. Confusion Matrix for expected agreement

		Observed		
Expected		Positive	Negative	Total
	Positive	(***a+b***)(***a+c***)/N	(***a+b***)(***b+d***)/N	***a+b***
	Negative	(***a+c***)(***c+d***)/N	(***c+d***)(***b+d***)/N	***c+d***
Total		***a+c***	***b+d***	N

$$expected\ agreement = \frac{\text{expected }(\boldsymbol{a}) + \text{expected }(\boldsymbol{d})}{\boldsymbol{N}} \tag{6}$$

Where, expected $(\boldsymbol{a}) = (\boldsymbol{a} + \boldsymbol{b})(\boldsymbol{a} + \boldsymbol{c})/\boldsymbol{N}$ and expected $(\boldsymbol{d}) = (\boldsymbol{c} + \boldsymbol{d})(\boldsymbol{b} + \boldsymbol{d})/\boldsymbol{N}$

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):

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

Where $\boldsymbol{a} = True\ Positives$ and $\boldsymbol{d} = Correct\ Negatives$.

Kappa Coefficient

Kappa is the measurement of the inter-rater reliability, which represents the extent to which the data collected in the study are correct representations of the variables measured [49]. The formula for kappa is:

$$Kappa = \frac{P_o - P_e}{1 - P_e} \tag{8}$$

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

Kappa coefficients are interpreted using the guidelines outlined by [50], 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 is used to measure the difference between the expected and the observed values from the environment that is being modeled [51]. The RMSE values can be used to distinguish model performance in a training period with that of a validation period as well as to compare the individual model performance to that of other predictive models. The RMSE of a model prediction for the estimated variable $\boldsymbol{X_{pred}}$ is defined as the square root of the mean squared error:

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

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 instance over negative instances [51]. It is a measure of the skill of different classifiers 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

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

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

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

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):

$$HR = \frac{No. of\ TP}{No. of\ Event} \quad (12)$$

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 [51].

Taylor Diagram

Taylor Diagram [52], provides a concise statistical summary and illustrates the matching patterns of data, for their Coefficients of Correlation, the root-mean-square, and their variances and ratio. Plots are patterned that precisely indicate measures and scores. The statistical display of the three factors on Taylor Diagram mathematically represented using the following formula:

$$E'^2 = \sigma_r^2 + \sigma_t^2 - 2\sigma_r\sigma_t\,\rho \tag{14}$$

Where; $\boldsymbol{\rho}$ = Correlation Coefficient

$\boldsymbol{E'}$ = Centered RMS difference between the observation and the prediction

$\boldsymbol{\sigma_r, \sigma_t}$= Variances of the observation and the prediction respectively

Results and Discussion

This section presents the analyzed data, and it constitutes the examination of how different climatic conditions influence the incidence of malaria. The comparisons between the two classifier methods were made for each of the 17 blocks in the district, and performance accuracy was evaluated month wise. The focus of this study, however, is to find the climatic influence of these incidents using some advanced machine learning techniques and ways to provide early warning about the possible future outbreaks.

Comparison of MLP and J48 Results

All three test options were used to assess the performance of both MLP and J48 methods, a) 10-fold cross-validation, and b) percent split (66%) and the user-supplied test sets. The following section discussed the outcome of the model prediction. An initial evaluation was completed over the Sundargarh district to investigate which of the machine learning method performed better compared to the others, before going for block-wise performance. All evaluations were done for the different months of the year to understand the prediction skill based on the monsoonal rainfall, temperature variation. While the results for prediction accuracy shows that the performance of both MLP and J48 is not very significantly different, but J48 shows better performance to MLP. In a similar study conducted by Gupta, Kumar & Sharma [53], where more attributes were analyzed and larger volumes of data used, the prediction using J48 has also turned out to be 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 shown in Figure 6 (Accuracy), Figure 7 (Kappa) and Figure 8 (RMSE), respectively.

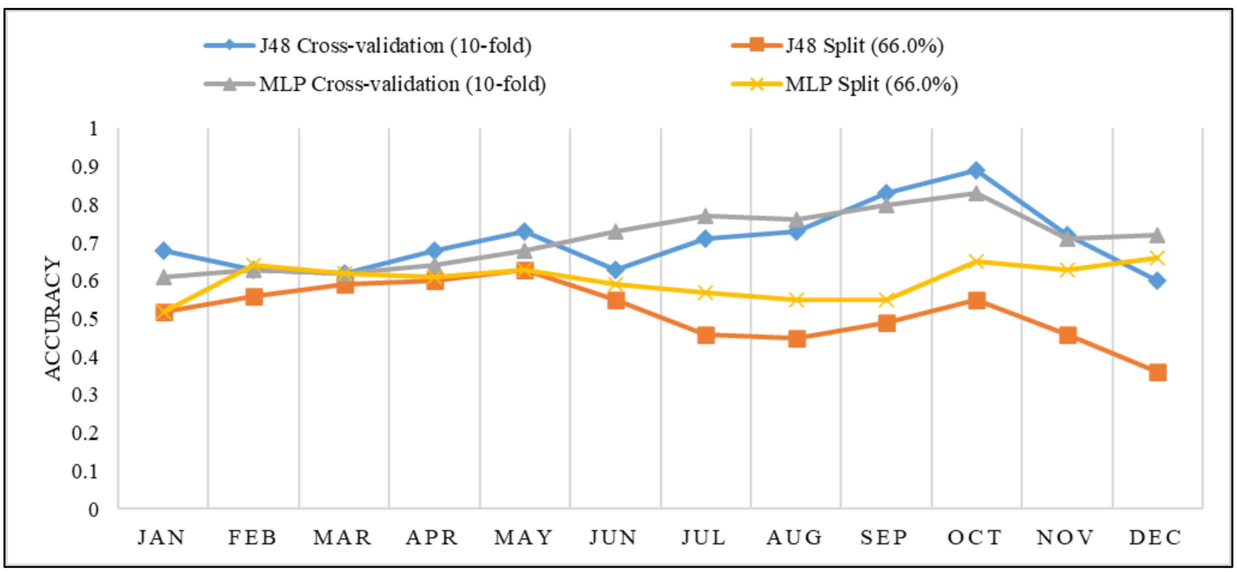


Figure 7. Month-wise Comparison of the Accuracy of the J48 (Cross-validation & Percent Split model) to the Multi-Layer Perceptron model for Accuracy for Sundargarh District

Figure 7 suggests that while 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 less accuracy.

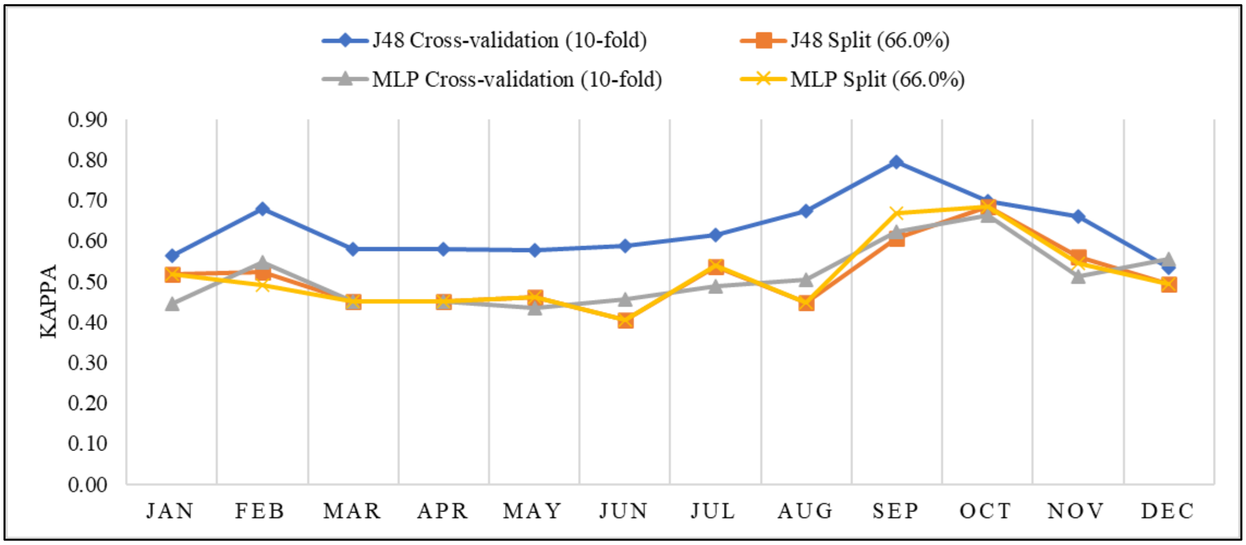


Figure 8. Month-wise Comparison of Kappa of the J48 (Cross-validation & Percent Split model) to the Multi-Layer Perceptron model for Accuracy for Sundargarh District

If we consider kappa, the J48 cross-validation method has significantly better kappa in comparison to the other three 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 Percent split (Figure 8).

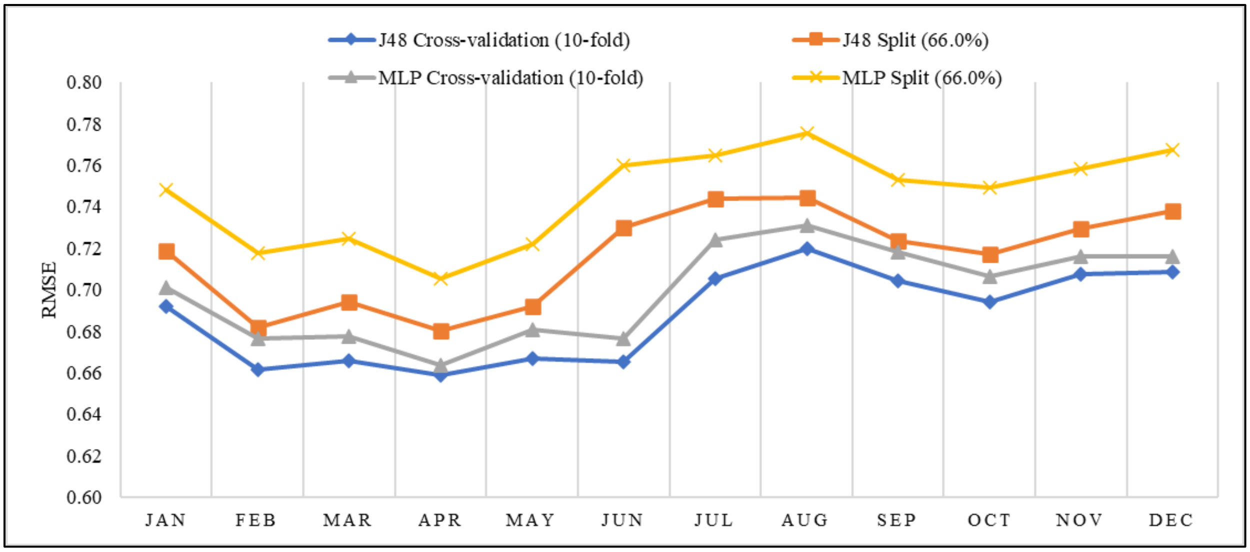


Figure 9. Month-wise Comparison of RMSE of the J48 (Cross-validation & Percent Split model) to the Multi-Layer Perceptron model for Accuracy for Sundargarh District

The prediction error (Figure 9) depicts that the cross-validation method has much less error compared to the percentage split method. The MLP percent split method consistently depicts the largest errors across seasons. Table 5 shows that the MLP has better accuracy, especially during the wet season, while less agreement to the observed condition as depicts comparatively lower kappa. For this, J48 has better performance during the wet season. Errors are more substantial for both the models during the wet period, and RMSE is low during the dry period. Since the data was analyzed monthly, J48 can be considered a more reliable predictor of malaria for the weather variables. During the monsoon and post-monsoon seasons, it has comparable RMSE and higher kappa (with highest values in September = 0.79 and October =0.70) indicated that it performed better compared to MLP with Kappa (September = 0.62 and October=0.66).

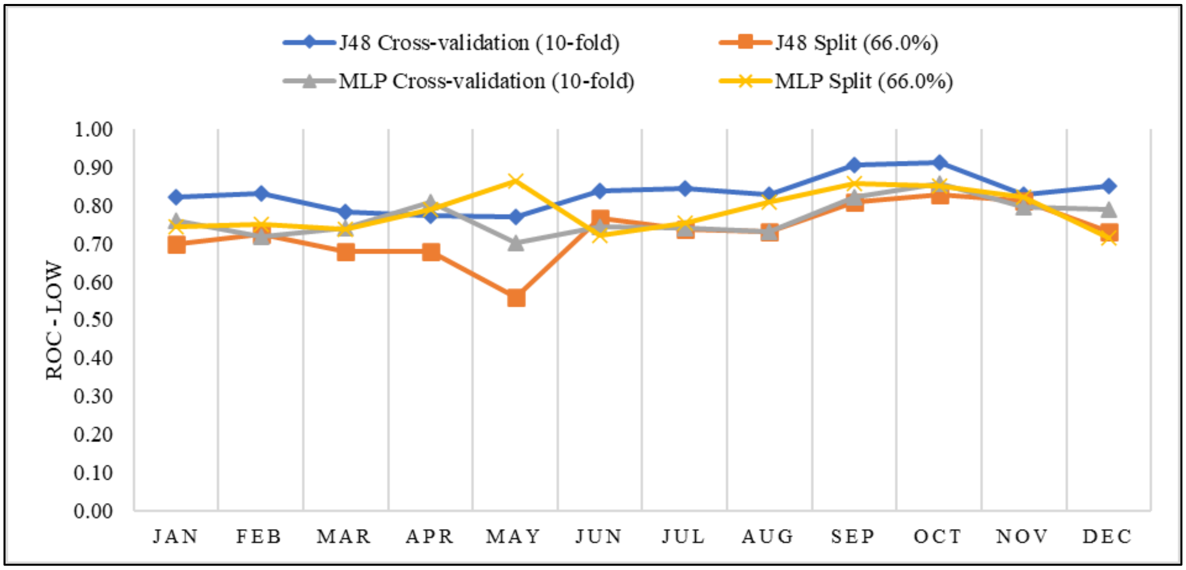
Table 5. Month-wise Performance metrics for RMSE, KAPPA, and Accuracy for the cross-validation Classifier

Month	Accuracy	Kappa	RMSE

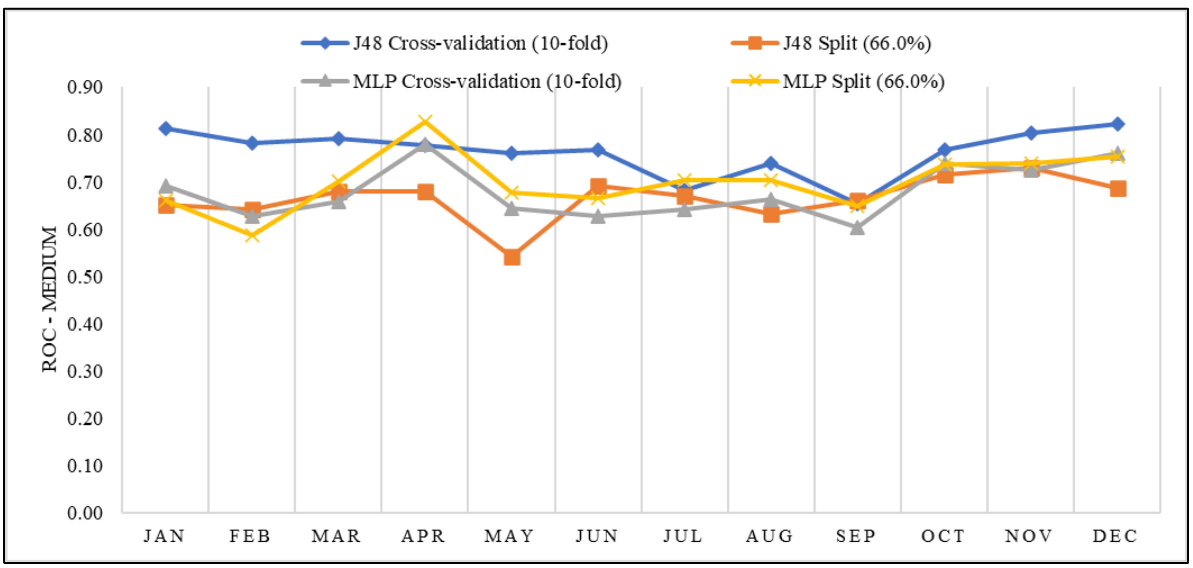
	J48	MLP	J48	MLP	J48	MLP
Jan	0.68	0.61	0.56	0.44	0.69	0.70
Feb	0.63	0.63	**0.68**	0.55	**0.66**	**0.68**
Mar	0.62	0.62	0.58	0.45	**0.67**	**0.68**
Apr	0.68	0.64	0.58	0.45	**0.66**	**0.66**
May	**0.73**	0.68	0.58	0.44	**0.67**	**0.68**
Jun	0.63	**0.73**	0.59	0.45	**0.67**	**0.68**
Jul	**0.71**	**0.77**	**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

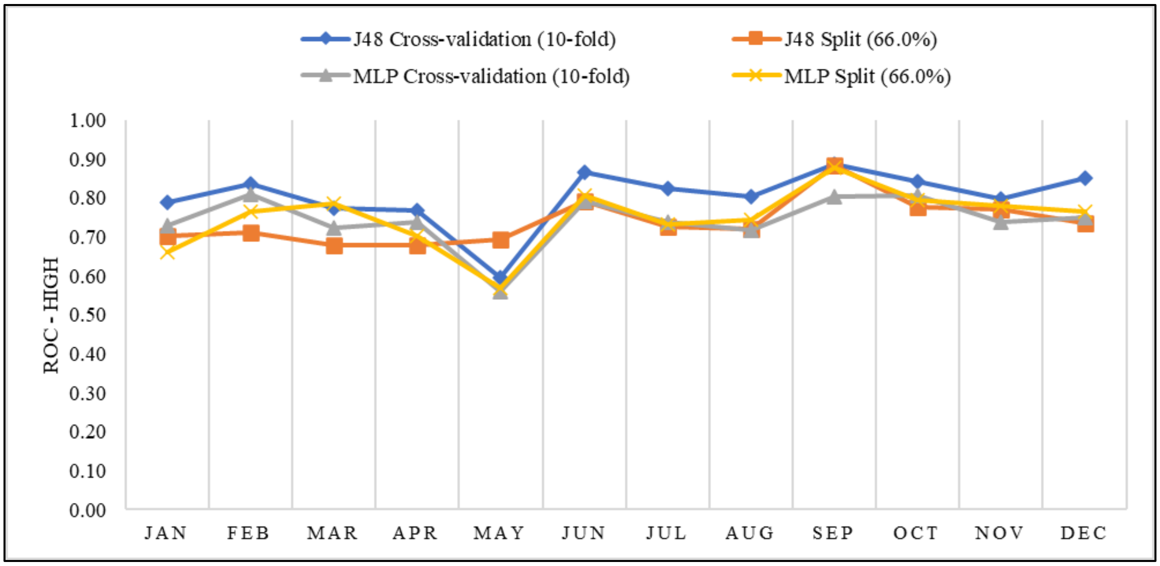
ROC score, as explained earlier, is generally a measure of the skill of the classifier. Evaluation with ROC requires the grouping of the prediction models into three distinct prediction categories, e.g., 1) High, 2) medium, and 3) Low. The evaluation shows how well the three categories of events can be predicted.


J48 Cross-validation (10-fold)
J48 Split (66.0%)
MLP Cross-validation (10-fold)
MLP Split (66.0%)
ROC - LOW
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC

(a)


J48 Cross-validation (10-fold)
J48 Split (66.0%)
MLP Cross-validation (10-fold)
MLP Split (66.0%)
ROC - MEDIUM
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC

(b)



(c)

Figure 10. Month-wise Comparison of ROC for ((a) Low, (b) medium, and (c) High prediction category) of the J48 (Cross-validation & Percent Split model) to the Multi-Layer Perceptron model for Accuracy for Sundargarh District

Figure 10 shows the performance of the three different event categories, and it is evident that the skill of the J48 method is comparatively better than MLP. Table 6 lists the ROC scores of all the four classifiers and provides a comparative analysis of the three-event categories.

Table 6. 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

The J48 cross-validation method has better performance in terms of predicting the 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. This probably could be because of the fewer sample datasets used in the Percent split methods for the training of the model, which might not be adequate. MLP has even poorer results depicting skill for high events with ROC=0.56 for May and consistently poor throughout the year. This shows that the models are generally poor during the early to the mid-monsoon period (May-August), irrespective of the model classifier technique used. At the same time, the prediction of the medium category event is challenging for both models as well. For J48 cross-validation, the lowest value (ROC=0.65) in September and for percent split method ROC=0.54 in May. While in the MLP cross-validation ROC=0.60 in September and ROC=0.59 in February month, respectively.

Performance Evaluation of the Models at smaller Administrative Units

The final exercise for the prediction model was to supply the classifiers with a user-defined set of datasets for training and isolate a specific year or years for prediction at block level. So, for this test, data from 2002 till 2015 was used for training and 2016 and 2017 for prediction. The results presented in Table 7 suggest the method has comparable performance to J48 and MLP. It has less RMSE and better accuracy and higher kappa values. Further investigating the performance of the Supplied test set, it was found 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). The block-wise comparison was presented in Table 7.

Table 7. Comparisons of Accuracy, RMSE, and Kappa for all blocks for J48 cross-validation and supplied set classifiers

Blocks	J48 Cross-Validation			J48 Supplied test set		
	Accuracy	Kappa	RMSE	Accuracy	Kappa	RMSE
Hemgiri	0.57	0.40	0.89	**0.71**	0.16	**0.37**
Lephripara	0.69	0.33	0.88	**0.83**	0.00	**0.34**
Tangarpali	**0.94**	**0.55**	0.76	**1.00**	**1.00**	**0.17**
Sundargarh	**0.88**	0.51	0.80	**1.00**	**1.00**	**0.19**
Subdega	**0.71**	**0.60**	0.86	**0.71**	0.26	**0.36**
Baragaon	**0.81**	**0.63**	0.81	**0.79**	0.17	**0.32**
Balisankara	0.55	0.31	0.89	**0.79**	0.00	**0.40**
Kutra	**0.98**	0.52	0.72	**1.00**	1.00	**0.16**
Rajgangpur	0.69	0.51	0.86	0.63	0.00	**0.39**
Kuanrmunda	0.63	**0.57**	0.87	0.63	0.30	**0.41**
Nuagaon	0.58	**0.55**	0.90	0.29	0.08	0.53
Bisra	0.53	0.38	0.89	0.58	0.39	**0.49**
Lathikata	**0.72**	0.50	0.86	0.42	0.13	**0.50**

Bonai	0.57	0.50	0.91	0.46	0.18	**0.48**
Lahunipara	**0.70**	**0.60**	0.86	0.54	0.47	**0.48**
Gurundia	0.53	0.45	0.90	0.17	0.07	0.58
Koira	0.51	0.39	0.90	0.17	0.29	0.58

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

ROC scores for both the classifiers of the J48 model were compared, and the results are presented in Table 8. The results 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. While blocks with plain land and forest cover performed much better compared to the highly elevated regions.

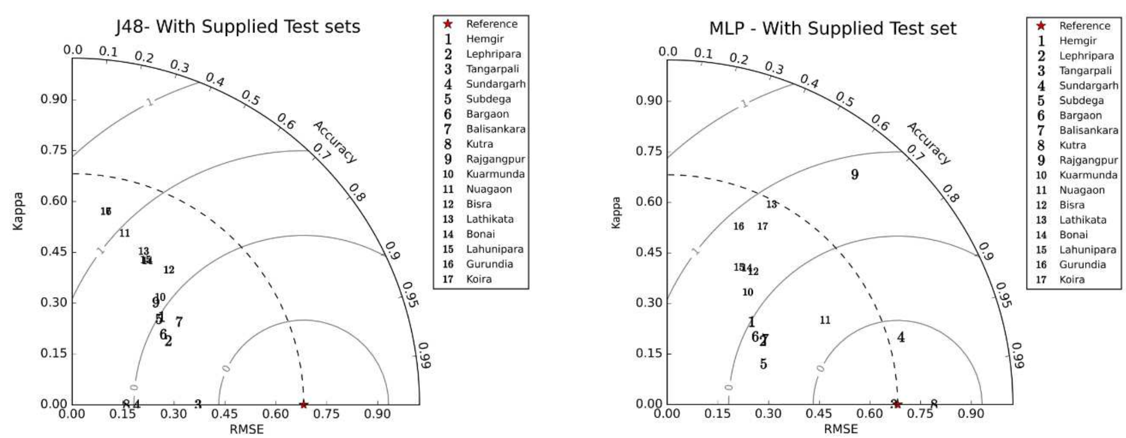
Table 8. Comparative analysis of ROC scores for all blocks in the District for J48 cross-validation and supplied set classifiers

Blocks	Cross-Validation			Supplied Test		
	High	Med	Low	High	Med	Low
Hemgiri	**0.83**	0.73	0.72	**0.76**	0.74	**0.94**
Lephripara	**0.75**	**0.76**	0.72	**0.79**	0.61	0.68
Tangarpali	0.69	**0.75**	**0.75**	**0.81**	**0.79**	**0.85**
Sundargarh	0.62	0.74	0.70	0.45	0.69	**0.81**
Subdega	**0.91**	**0.76**	**0.86**	**0.98**	**0.84**	**0.93**
Baragaon	**0.93**	**0.77**	**0.90**	0.73	**0.76**	0.60
Balisankara	0.73	0.69	0.74	0.12	0.34	0.64
Kutra	**0.99**	**0.80**	**0.90**	**0.86**	**0.94**	**0.88**

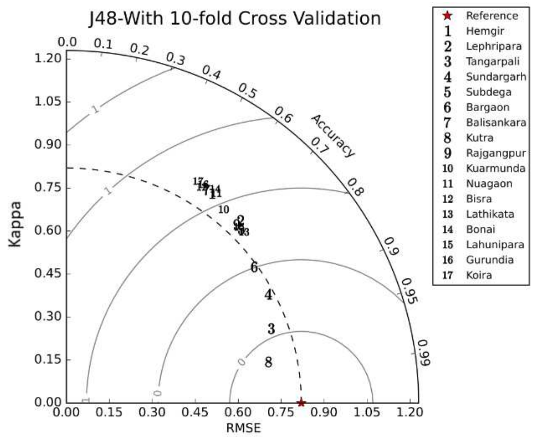
Rajgangpur	**0.82**	**0.76**	**0.84**	**0.92**	**0.89**	**0.77**
Kuanrmunda	**0.97**	0.71	**0.84**	0.93	0.65	0.83
Nuagaon	**0.81**	0.67	**0.81**	0.60	0.71	0.53
Bisra	**0.88**	0.70	**0.84**	0.49	**0.79**	0.73
Lathikata	**0.88**	**0.89**	0.66	0.65	**0.87**	0.43
Bonai	**0.76**	0.63	**0.78**	0.46	**0.82**	0.59
Lahunipara	**0.90**	**0.76**	**0.94**	0.31	0.39	0.39
Gurundia	**0.89**	0.71	**0.76**	0.53	0.20	0.33
Koira	**0.77**	**0.75**	0.72	0.25	0.21	0.34

Highlighted values are ROC scores ≥ 0.75

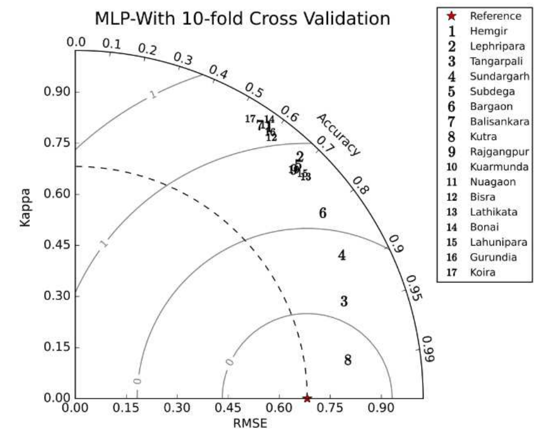
Figure 11 demonstrates the comparison of the three test options used in both classifier techniques J48 and MLP using the Taylor diagram. All three indicators (Accuracy, Kappa, and RMSE) were represented in three different axes. RMSE in X-axis, Kappa in Y-axis, and accuracy in the arc, respectively.



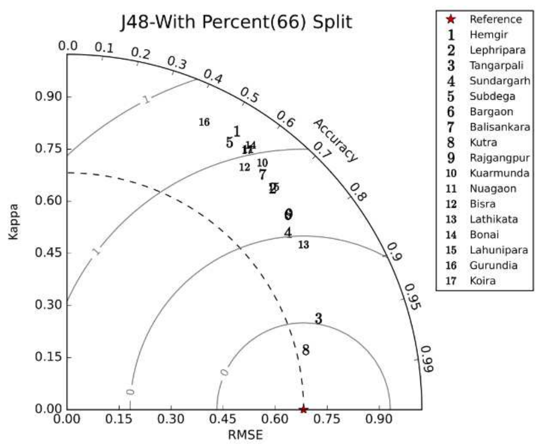
(a) (b)



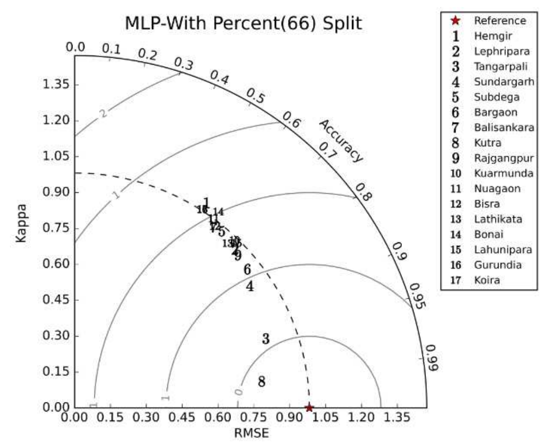
(c)



(d)



(e)



(f)

Figure 11. Block-wise accuracy (arc), RMSE (x-axis) and Kappa (Y-axis) for the J48 and the Multi-Layer Perceptron model with a Supplied test set, 10-fold Cross-validation & Percent Split classifiers

While the numbers 1 to 17 represent each block of the district, its position in the plotting space determines its corresponding Accuracy, RMSE, and Kappa. Block-wise analysis with all the results suggests that with the 10-fold cross-validation and the supplied test set option has yielded promising results in comparison to the percent split and supplied test options. Especially blocks (8=Kutra, 3=Tangarapali, 4=Sundargarh) from the central to western plain have better performance to the blocks with varying topography (17=Koira, 1=Hemgiri, 14=Bonai). The supplied test options depicted smaller RMSEs but also have inconsistency with Accuracy and Kappa (either Accuracy=1.0 or very low), making it unreliable for use in predictions. So, it is evident that flat terrains with lower variability in the rainfall, temperature, and humidity provides reliable performance than to the regions with higher variability.

Month-wise Nominal Relationship

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 derived from the continuous numeric data range [54] is represented in Table 9. Based on the spectrum, Table 10 provides a summary of the relationship between the nominal rainfall, temperature, and humidity range to that of the malaria incidents.

Table 9. Climate and Incident data range Evaluation

Range	RF	T2max	RH	Incidents
Low	0 - 23	28.3 - 30.3	68.3 - 82.6	35 - 78
Medium	23.1 - 178	30.4 - 33.6	82.7 - 93	78.1 - 173
High	178.1 - 445	33.7 - 38.3	93.1 - 96.7	173.1 - 460
Very High	> 445	> 38.3	> 96.7	> 460

Table 10. Month-wise nominal relationship between incident data and climate data

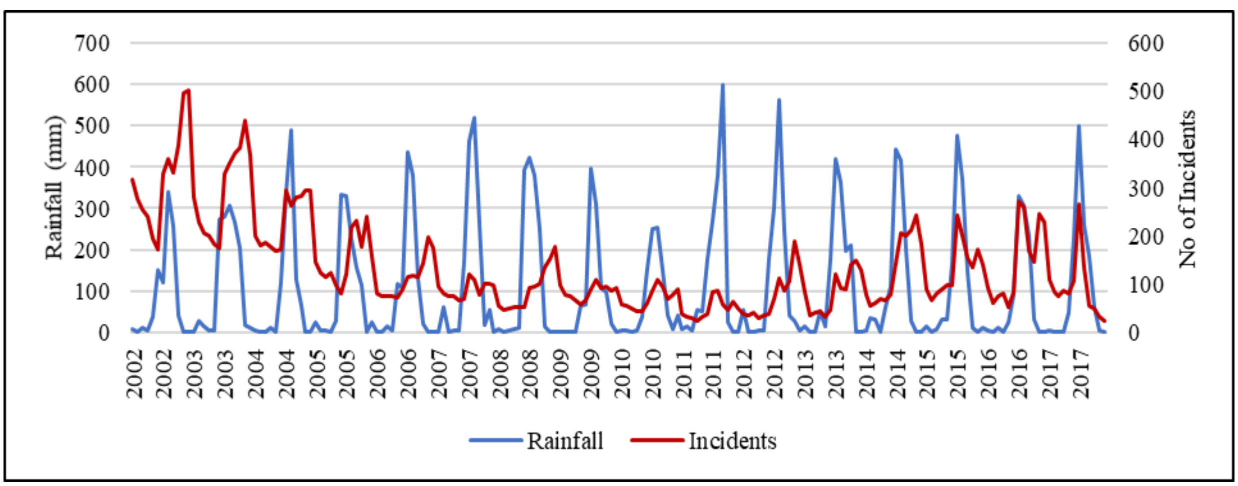
Month	RF	$T2_{max}$	RH	Incidents
January	Low	Low	Low	Low
February	Low	Low	Low	Low
March	Low	Medium	Low	Medium
April	Low	Very High	Low	Medium

May	Medium	Very High	Medium	Medium
June	High	High	Medium	Medium
July	High	Medium	High	High
August	Very High	Medium	High	High
September	Very High	Low	High	Medium
October	Medium	Low	Medium	Medium
November	Low	Low	Medium	High
December	Low	Low	Low	High

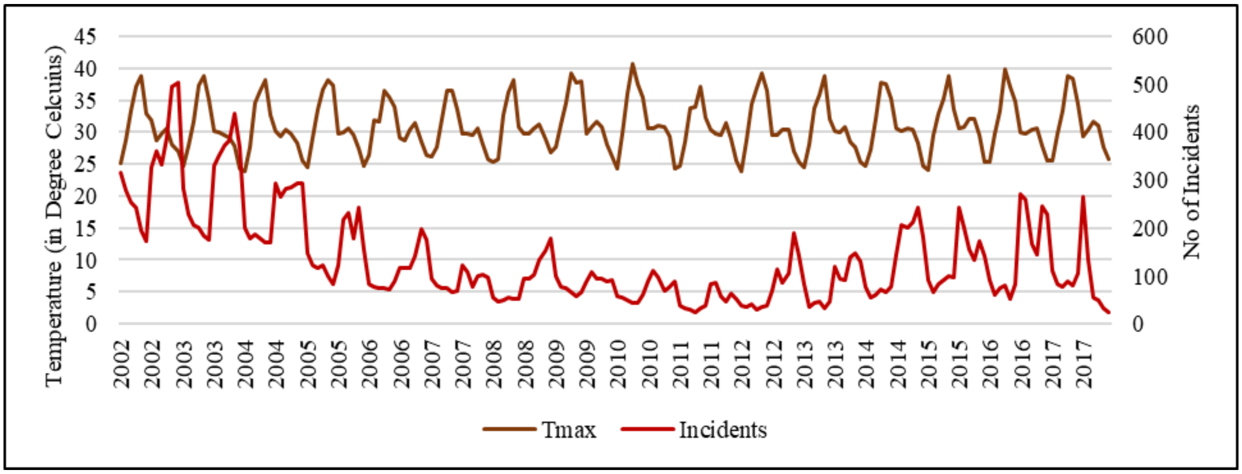
As shown in the table, periods of low temperature (28°-30°), low rainfall (0 -23mm), and low relative humidity (68-82%) during the drier and cooler months of January and February are characterized by lower cases of malaria. During the months of March-April-May, a period of low precipitation, medium to higher temperature with lower relative humidity, there is an increase in the number of incidences of malaria. This also agrees with the findings of Lee et al. [55], 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, higher temperature, and medium to high relative humidity, the malaria incidents were further on the rise. Surprisingly, malaria incidents climbed even after the withdrawal of the monsoon, fall in the temperature and humidity significantly during the November-December winter period. So, there is possibly a lag-effect of the climatic phenomenon on the incidents.

Based on the results, it can be concluded that relative humidity and temperature showed a strong association with malaria incidence, which is consistent with the study by Srimath-Tirumula [56], in Vishakhapatnam, in India, which experiences similar climate compared to Sundargarh. In contrast, rainfall showed a relatively weaker association, which is in line with the study by Bomblies [3], which argues that during the rainy season, the breeding habitats of mosquitoes are flushed away temporarily. Still, they start breeding again when the rains stop, and water becomes stagnant, and the environmental condition is conducive for breading.

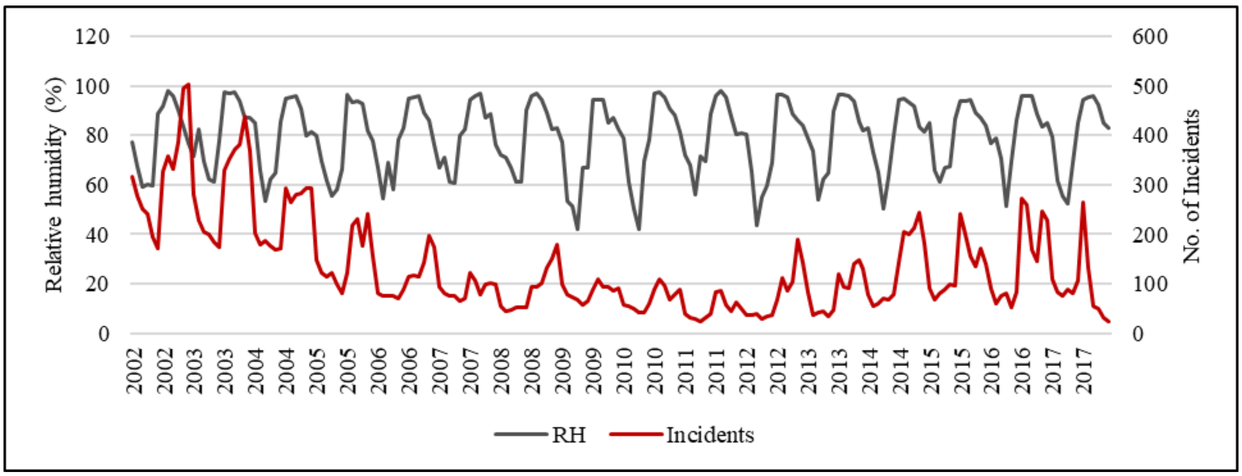
This study found that extremely high temperature is one of the crucial triggers of the higher number of malaria incidents in Sundargarh district. Therefore, this agrees with the argument of Smith et al. [57], when the temperatures increase, it reduces the time taken by the mosquito parasite to complete its development. Furthermore, relative humidity also affects the transmission of the malaria vector in agreement with [58], found out that mosquitoes survive better under high humidity conditions. During high humid seasons, the number of malaria incidents increases compared to the less humid conditions. The time series plots (Figure 12) for the rainfall (a), temperature (b), and humidity (c) submit its direct association with the incidents reported.



(a)



(b)



(c)

Figure 12. Time series plot for Monthly Rainfall (a), Maximum Temperature (b), and Relative Humidity (c) in comparison with malaria incidents for the period of 2002-2017

It is concluded that relative humidity and temperature showed a significant relationship to malaria incidences in the district, especially for some blocks, those are in flat terrain and near dense vegetation like the forest. Additionally, rainfall also affects the transmission of malaria vector incidence. However, the rate of vector transmission during the rainy season is relatively lower, suggesting the influence of rain on malaria incidents may happen in a time lag mode as well.

Discussions

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 [59-61]. Several such attempts were proposed, such as a spatial decision support system for Karnataka [62] or the operational system in Kenya [60], but not all are very successful and useful. Statistical regression-based analysis like the multiple polynomial regression, semiparametric Poisson distribution methods, distributed non-linear lag model, hierarchical Bayesian framework and the time series regression models [14-19, 20-22, 56], or use of ANN or machine learning as discussed by researchers, provides an opportunity to process the dataset and establish an association between climate and the malaria incidents. Considering the climate dimension only in malaria early warning is not adequate and requires a deep understanding of the influence of all other facets, including climate in the establishment of an effective and operational warning system.

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 [63]. Since we are typically 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 [63]. Working with small sample sizes in machine learning is acceptable but choosing the correct sampling method is vital [64, 65]. 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 would yield better results [65].

Assessments made by researchers to evaluate different classifier performances [66] recommended the use of leave-one-out cross-validation method as a preferred method of prediction. Thus, this study put an effort to assess the 10-fold cross-validation method, which incidentally also performed well. With more in-depth analysis, the reason would be linked to the data sampling and training strategy implemented by the method in comparison 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 that arises 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, a process by which the parameters were tuned with an independent validation set, that suggested the ideal number of trees in a classifier, hidden layer size (activation function) in the Neural network.

The probable explanation for the model not performing well for some of the blocks and months could be a factor that is external to the climate influence. Furthermore, the interventions in place in parts of the district and other regions [67] and the socio-economic status of the significant population in the eastern belt, being tribe and access to necessary facilities is limited as also explained by Sundararajan et al. [13]. These factors influence the increase or decrease in the cases, and not truly reflect the direct influence of climate. The other reflection from the analysis is that the model's performance dips down significantly, especially during July, August. It picks up during September October and then again shows lower accuracy during November and December months (this is depicted in Figure 7). The poor performance of the model could be because of the varying topography, which affects the intra-seasonal rainfall variability as well as the spatial variation of the temperature and humidity could influence the results strongly. For example, blocks such as Bonai, Koira, Gurundia, and Hemgiri are with higher elevation and depict considerably lower accuracy for the prediction. In comparison, blocks with plain land and forest cover had better performance.

Conclusion

The climatic condition of Odisha, especially the Sundargarh district, makes it vulnerable to malaria y. Monsoon rainfall, maximum surface temperature ranging from 27° to 40° Celsius during the summer, and relative humidity in the range of 60% to 85% provide a more favorable climatic condition for the breeding of the malaria larva during the monsoon and post-monsoon period. It was found that the increase in malaria incidents is significantly attributed to climatic factors such as temperature, humidity, and monthly rainfall variability. Among the two classifier models used, J48 has shown comparatively better skills over the MLP. J48 demonstrated less error (RMSE = 0.6), better kappa = 0.63, and higher accuracy = 0.71), suggesting it to be a suitable model.

J48 model has provided a greater insight into the predictability of malaria when compared in seasonal scale. During the pre-monsoon (Mar-Apr-May) period it has accuracy = 0.68, kappa = 0.58 and RMSE = 0.67, for monsoon (Jun-July-Aug-Sep) period with accuracy = 0.73, kappa = 0.67 and RMSE = 0.70, post-monsoon (Oct-Nov) period with highest accuracy = 0.81, kappa = 0.68 and RMSE = 0.70 and during the winter (Dec-Jan-Feb) period with lowest accuracy = 0.64, kappa = 0.59 and RMSE = 0.69, respectively. This suggests that the model performance is particularly good during the monsoon and post-monsoon when the malaria incidents are at the peak. Besides, non-climatic factors play a significant role in the malaria spreading, which was reflected with a lower accuracy. However, climate being an extremely complex and variable factor to predict, the results provided promising signal for the prediction of future malaria incidents. Therefore, it is recommended that the public health department could adopt the J48 classifier machine learning technique in the malaria early warning system for the early detection of malaria.

Even though the models have shown better performance in terms of predicting malaria incidence, it is constrained by the non-availability of datasets for an elongated period. More finer scale datasets (both climate and malaria cases) would have provided an opportunity for deeper analysis to understand the phases and lags within a month as well. Furthermore, non-climatic factors such as the demography, immunity within the population, the socio-economic structure of society, 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.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or 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.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Author Contributions

PM and NKT conceptualized the study and drafted the manuscript. PM prepared the data set for analysis. NKT contributed to the study design. NKT, IP, and SS provided critical input and improvising the manuscript. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

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

Conflict of interest disclosure

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

References

1. Organization WH. World malaria report 2018. 2018. World Health Organization: Geneva. 2020.
2. Kovats RS, Bouma MJ, Hajat S, Worrall E, Haines A. El Niño and health. The Lancet. 2003; 362(9394):1481-9.
3. Bomblies A. Modeling the role of rainfall patterns in seasonal malaria transmission. Climatic change. 2012;112(3-4):673-85.
4. Githeko AK, Lindsay SW, Confalonieri UE, Patz JA. Climate change and vector-borne diseases: a regional analysis. Bulletin of the World Health Organization. 2000;78:1136-47.
5. Leal Filho W, Bönecke J, Spielmann H, Azeiteiro UM, Alves F, de Carvalho ML, et al. Climate change and health: An analysis of causal relations on the spread of vector-borne diseases in Brazil. Journal of Cleaner Production. 2018;177:589-96.
6. Van Lieshout M, Kovats R, Livermore M, Martens P. Climate change and malaria: analysis of the SRES climate and socio-economic scenarios. Global environmental change. 2004;14(1):87-99.
7. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, et al. The genome sequence of the malaria mosquito Anopheles gambiae. science. 2002;298(5591):129-49.
8. Kakmeni FMM, Guimapi RY, Ndjomatchoua FT, Pedro SA, Mutunga J, Tonnang HE. Spatial panorama of malaria prevalence in Africa under climate change and interventions scenarios. International Journal of Health Geographics. 2018;17(1):1-13.

9. Programme NVBDC. Malaria situation in India. Ministry of Health & Family Welfare, Govt. of India: Directorate General of Health Services; [Available from: https://nvbdcp.gov.in/index4.php?lang=1&level=0&linkid=564&lid=3867.10. Accessed 28 March 2020.

10. Organization WH. World malaria report 2019. World Health Organization: Geneva. 2019.

11. Mahakur PK, Nayak NC. Intrastate Income Inequalities in Odisha: Examining Decomposition by Regions and Broad Sectors. Odisha Economy Discussion Series. 2019;1.

12. Kannan KP. Interrogating inclusive growth: Poverty and inequality in India. Routledge; 2017 Sep 19

13. Sundararajan R, Kalkonde Y, Gokhale C, Greenough PG, Bang A. Barriers to malaria control among marginalized tribal communities: a qualitative study. PloS One. 2013;8(12):e81966.

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

15. Shimaponda-Mataa NM, Tembo-Mwase E, Gebreslasie M, Achia TN, Mukaratirwa S. Modelling the influence of temperature and rainfall on malaria incidence in four endemic provinces of Zambia using semiparametric Poisson regression. Acta tropica. 2017;166:81-91.

16. Guo C, Yang L, Ou C-Q, Li L, Zhuang Y, Yang J, et al. Malaria incidence from 2005–2013 and its associations with meteorological factors in Guangdong, China. Malaria Journal. 2015;14(1):116.

17. Arab A, Jackson MC, Kongoli C. Modelling the effects of weather and climate on malaria distributions in West Africa. Malaria Journal. 2014;13(1):126.

18. Imai C, Armstrong B, Chalabi Z, Mangtani P, Hashizume M. Time series regression model for infectious disease and weather. Environmental Research. 2015;142:319-27.

19. Rejeki DSS, Nurhayati N, Budi A, Murhandarwati EEH, KUSNANTO H. A Time Series Analysis: Weather Factors, Human Migration and Malaria Cases in Endemic Area of Purworejo, Indonesia, 2005–2014. Iranian Journal of Public Health. 2018;47(4):499.

20. Neter J, Kutner MH, Nachtsheim C, Wasserman W. Applied linear statistical models. 1996. WCB McGraw-Hill. 1996.

21. Basheer IA, Hajmeer M. Artificial neural networks: fundamentals, computing, design, and application. Journal of Microbiological Methods. 2000;43(1):3-31.

22. Yao J, Tan CL, Poh H-L. Neural networks for technical analysis: a study on KLCI. International Journal of Theoretical and Applied Finance. 1999;2(02):221-41.

23. OPERATIONS DOC, ODISHA. District Census Handbook Sundargarh. 2011; SERIES-22. [Available from: https://censusindia.gov.in/2011census/dchb/2105_PART_B_DCHB_SUNDARGARH.pdf]. Accessed 17 April 2020.

24. Servadio JL, Rosenthal SR, Carlson L, Bauer C. Climate patterns and mosquito-borne disease outbreaks in South and Southeast Asia. Journal of Infection and Public Health. 2018;11(4):566-71.

25. Jolivet R, Grandin R, Lasserre C, Doin MP, Peltzer G. Systematic InSAR tropospheric phase delay corrections from global meteorological reanalysis data. Geophysical Research Letters. 2011;38(17).

26. Marcos R, González-Reviriego N, Torralba V, Soret A, Doblas-Reyes FJ. Characterization of the near surface wind speed distribution at global scale: ERA-Interim reanalysis and ECMWF seasonal forecasting system 4. Climate Dynamics. 2019;52(5-6):3307-19.

27. Parker WS. Reanalyses and observations: What's the difference? Bulletin of the American Meteorological Society. 2016;97(9):1565-72.

28. Bengtsson L, Hagemann S, Hodges KI. Can climate trends be calculated from reanalysis data? Journal of Geophysical Research: Atmospheres. 2004;109(D11).

29. Holmes G, Donkin A, Witten IH. Weka: A machine learning workbench. In Proceedings of ANZIIS'94-Australian New Zealnd Intelligent Information Systems Conference 1994 Nov 29 (pp. 357-361). IEEE.

30. Witten IH, Frank E. Data mining: practical machine learning tools and techniques with Java implementations. Acm Sigmod Record. 2002 Mar 1;31(1):76-7.

31. Hornick M, inventor; Oracle International Corp, assignee. Data mining agents for efficient hardware utilization. United States patent US 7,620,635. 2009 Nov 17.

32. Bui QT, Nguyen QH, Pham VM, Pham MH, Tran AT. Understanding spatial variations of malaria in Vietnam using remotely sensed data integrated into GIS and machine learning classifiers. Geocarto International. 2019 Oct 15;34(12):1300-14.

33. Sharma V, Kumar A, Lakshmi Panat D, Karajkhede G. Malaria outbreak prediction model using machine learning. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET). 2015 Dec;4(12).

34. Olayinka TC, Chiemeke SC. Predicting Paediatric Malaria Occurrence Using Classification Algorithm in Data Mining. Journal of Advances in Mathematics and Computer Science. 2019 Apr 1:1-0.

35. Mello-Román JD, Mello-Román JC, Gomez-Guerrero S, García-Torres M. Predictive Models for the Medical Diagnosis of Dengue: A Case Study in Paraguay. Computational and mathematical Methods in Medicine. 2019 Jul 29;2019.

36. Guo P, Liu T, Zhang Q, Wang L, Xiao J, Zhang Q, Luo G, Li Z, He J, Zhang Y, Ma W. Developing a dengue forecast model using machine learning: A case study in China. PLoS neglected tropical diseases. 2017 Oct 16;11(10):e0005973.

37. Atulbhai DK. COMPARISION AND COMBINATION OF MINING TECHNIQUES FOR GENE ANALYSIS TO IDENTIFY DENGUE.

38. Shakil KA, Anis S, Alam M. Dengue disease prediction using weka data mining tool. arXiv preprint arXiv:1502.05167. 2015 Feb 18.

39. Leo J, Luhanga E, Michael K. Machine Learning Model for Imbalanced Cholera Dataset in Tanzania. The Scientific World Journal. 2019 Aug 1;2019.

40. Mahmud SH, Hossin MA, Ahmed MR, Noori SR, Sarkar MN. Machine Learning Based Unified Framework for Diabetes Prediction. InProceedings of the 2018 International Conference on Big Data Engineering and Technology 2018 Aug 25 (pp. 46-50).

41. Zia UA, Khan N. Predicting diabetes in medical datasets using machine learning techniques. International Journal of Scientific & Engineering Research Volume. 2017 May;8(5).

42. Al Jarullah AA, editor Decision tree discovery for the diagnosis of type II diabetes. 2011 International Conference on Innovations in Information Technology; 2011: IEEE.

43. Sabarinathan V, Sugumaran V. Diagnosis of heart disease using decision tree. International Journal of Research in Computer Applications & Information Technology. 2014;2(6):74-9.

44. Dangare CS, Apte SS. Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications. 2012;47(10):44-8.

45. Korting TS. C4.5 algorithm and multivariate decision trees. Image Processing Division, National Institute for Space Research–INPE Sao Jose dos Campos–SP, Brazil. 2006.

46. Nicholson C. A Beginner's Guide to Multilayer Perceptrons (MLP) [Available from: https://pathmind.com/wiki/multilayer-perceptron#three.%20https://pathmind.com/wiki/multilayer-perceptron#three]. Accessed 20 April 2020.

47. Quinlan JR. C4.5: programs for machine learning: Elsevier; 2014.

48. Kaur G, Chhabra A. Improved J48 classification algorithm for the prediction of diabetes. International Journal of Computer Applications. 2014;98(22).

49. McHugh ML. Interrater reliability: the kappa statistic. Biochemia medica: Biochemia medica. 2012;22(3):276-82.

50. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics. 1977:363-74.

51. Kumar N, Khatri S, editors. Implementing WEKA for medical data classification and early disease prediction. 2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT); 2017: IEEE.

52. Taylor KE. Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research: Atmospheres. 2001;106(D7):7183-92.

53. Gupta S, Kumar D, Sharma A. Performance analysis of various data mining classification techniques on healthcare data. International Journal of Computer Science & Information Technology (IJCSIT). 2011 Aug;3(4):155-69.

54. Mangiafico S. Summary and analysis of extension program evaluation in R, version 1.15. 0. Rutgers Cooperative Extension, New Brunswick, NJ: https://rcompanion org/handbook/[Google Scholar]. 2016.

55. Lee E, Burkhart J, Olson S, Billings AA, Patz JA, Harner EJ. Relationships of climate and irrigation factors with malaria parasite incidences in two climatically dissimilar regions in India. Journal of Arid Environments. 2016 Jan 1;124:214-24.

56. Srimath-Tirumula-Peddinti RCPK, Neelapu NRR, Sidagam N. Association of climatic variability, vector population and malarial disease in district of Visakhapatnam, India: a modeling and prediction analysis. PLoS One. 2015;10(6):e0128377.

57. Smith DL, Perkins TA, Tusting LS, Scott TW, Lindsay SW. Mosquito population regulation and larval source management in heterogeneous environments. PloS One. 2013;8(8):e71247.

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

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

60. Thomson M, Indeje M, Connor S, Dilley M, Ward N. Malaria early warning in Kenya and seasonal climate forecasts. The Lancet. 2003;362(9383):580.

61. Thomson MC, Connor SJ. The development of malaria early warning systems for Africa. Trends in parasitology. 2001;17(9):438-45.

62. Shekhar S, Yoo E, Ahmed S, Haining R, Kadannolly S. Analysing malaria incidence at the small area level for developing a spatial decision support system: a case study in Kalaburagi, Karnataka, India. Spatial and spatio-temporal epidemiology. 2017;20:9-25.

63. Raschka S. Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:181112808. 2018.

64. Lusa L. Joint use of over-and under-sampling techniques and cross-validation for the development and assessment of prediction models. BMC Bioinformatics. 2015;16(1):1-10.

65. Cawley GC, Talbot NL. On over-fitting in model selection and subsequent selection bias in performance evaluation. The Journal of Machine Learning Research. 2010;11:2079-107.

66. Bischl B, Mersmann O, Trautmann H, Weihs C. Resampling methods for meta-model validation with recommendations for evolutionary computation. Evolutionary Computation. 2012 Jun;20(2):249-75.

67. Sahu S, Gunasekaran K, Raju H, Vanamail P, Pradhan M, Jambulingam P. Response of malaria vectors to conventional insecticides in the southern districts of Odisha State, India. The Indian Journal of Medical Research. 2014;139(2):294.

Figures

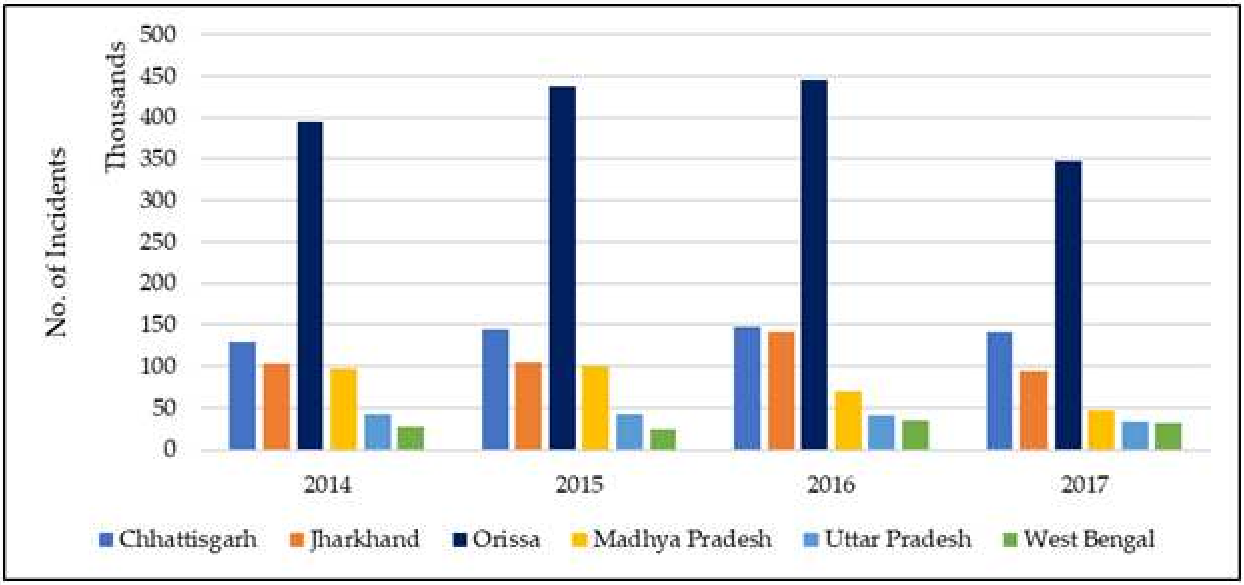


Figure 1

Malaria incidents across six different states of India for the period of 2014-2017.

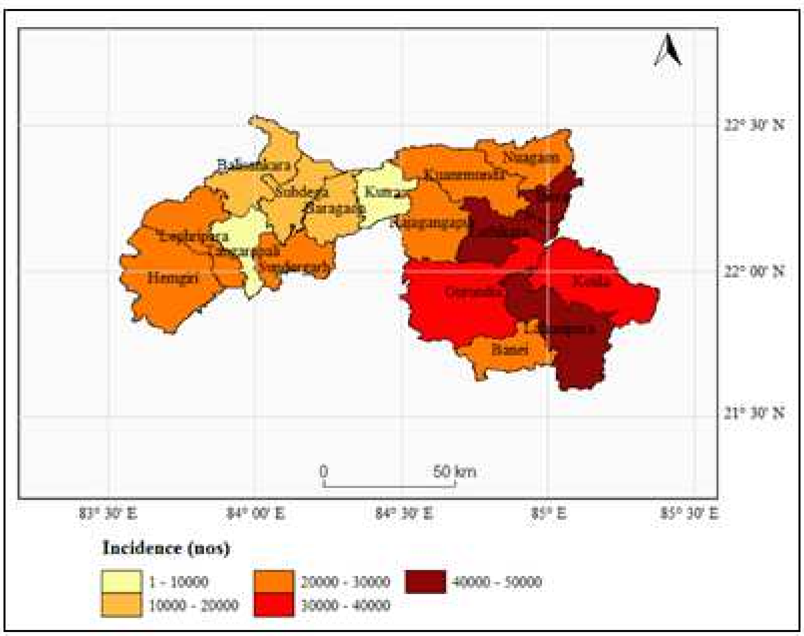


Figure 2

Cumulative malaria incidence map for the period 2002 to 2017 (Data Source: Directorate of Public Health, Odisha)

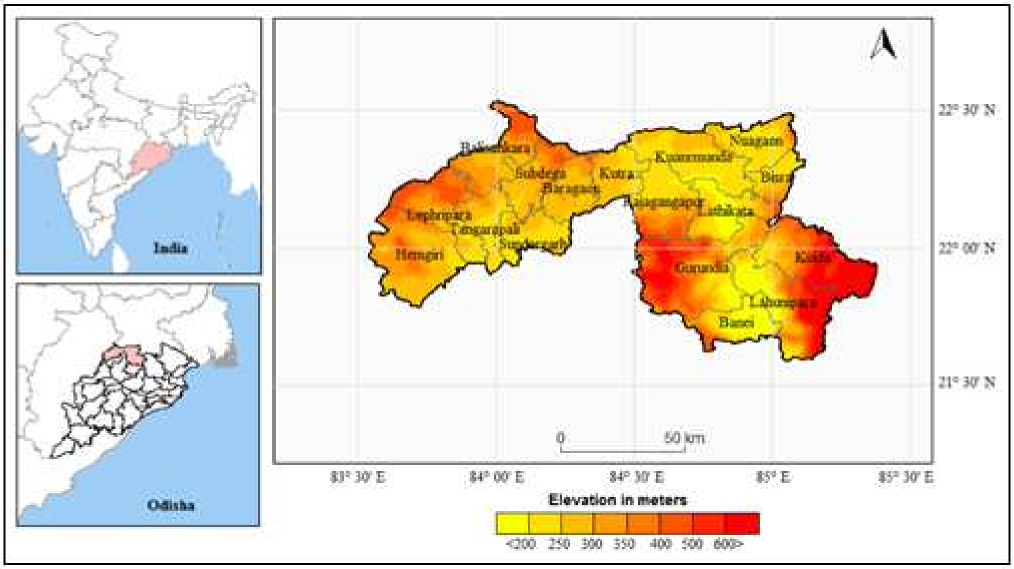


Figure 3

Study area with elevation for Sundargarh district in Odisha, India

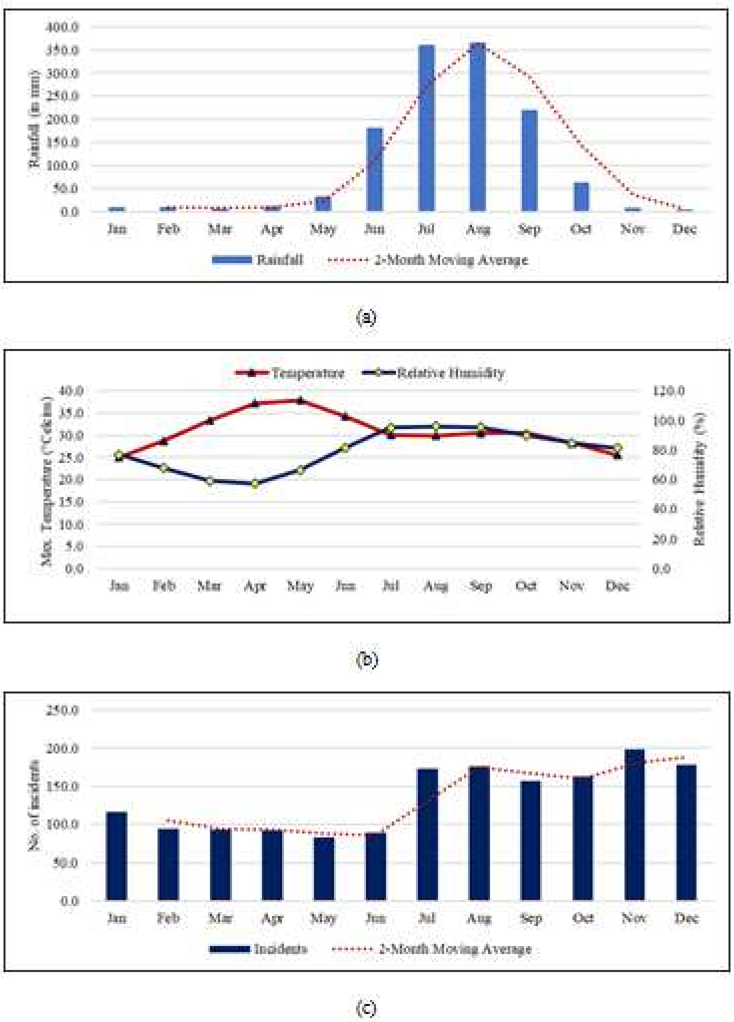
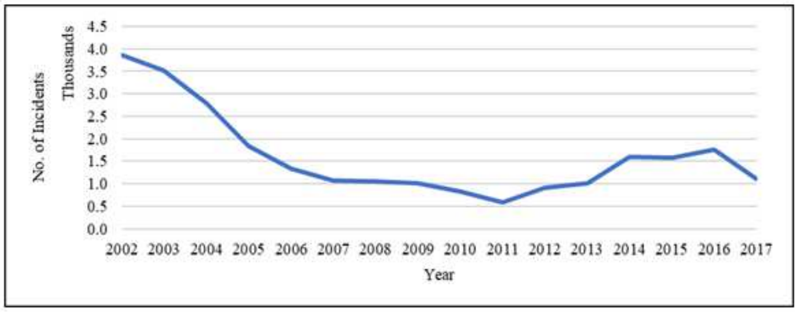
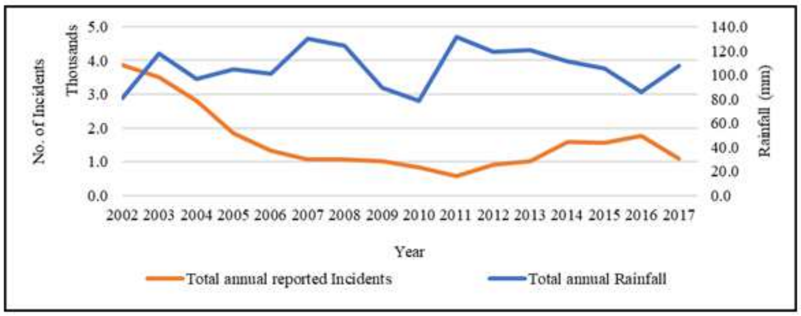


Figure 4

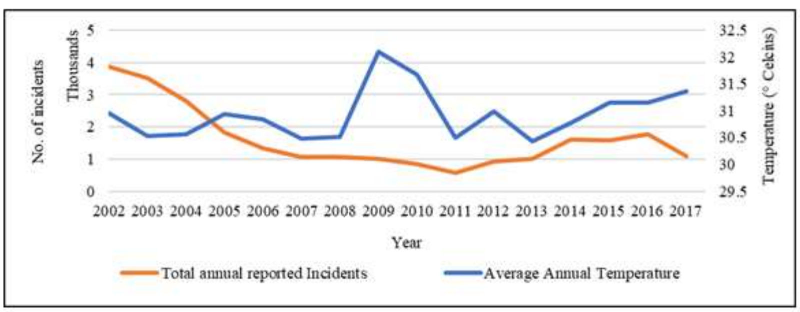
Average monthly rainfall (in mm) (a), maximum surface temperature (in °C), relative humidity (in %) (b), and incidents reported (c) for the period of 2002 to 2017.



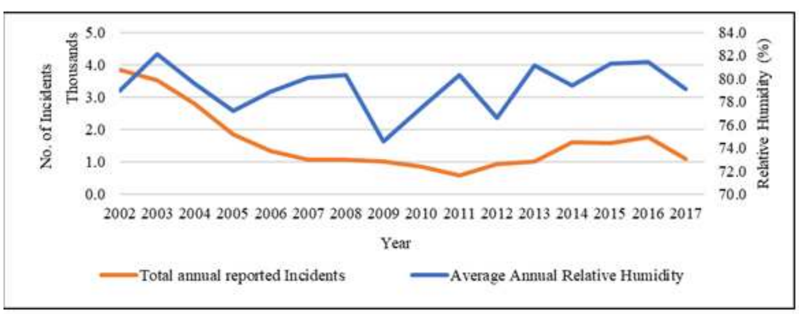
(a)



(b)



(c)



(d)

Figure 5

The trend of annual incidents (a); and comparison with average annual rainfall (b); temperature (c); and relative humidity (d) for the period of 2002 to 2017

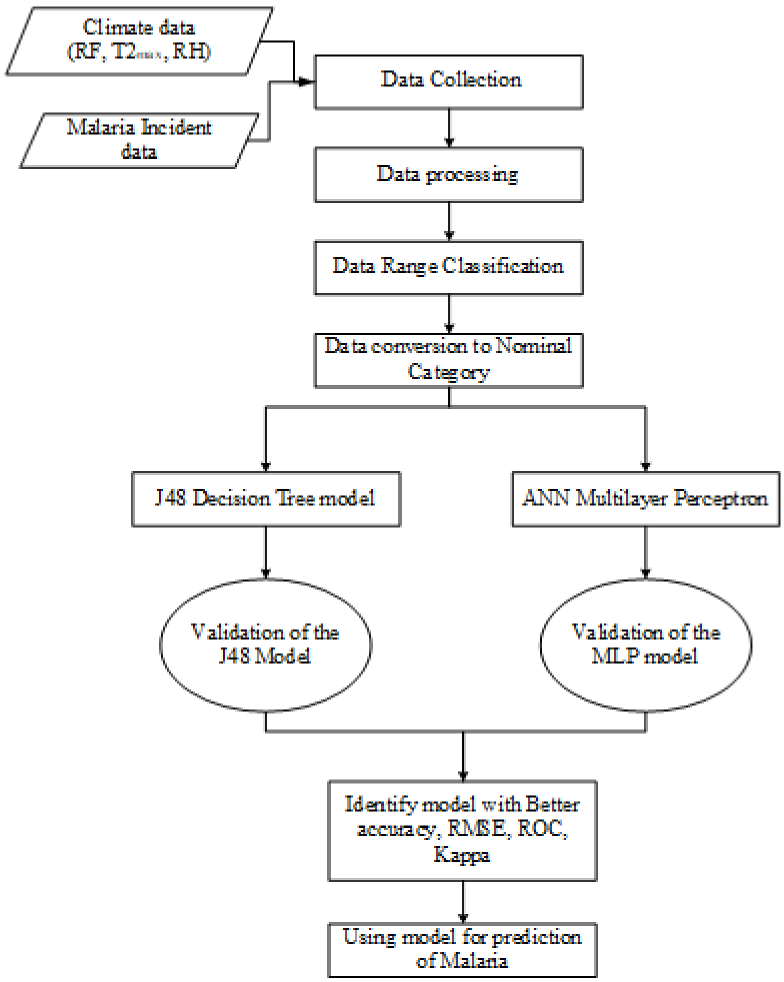


Figure 6

Detailed Methodology for Predictive Modeling of Malaria

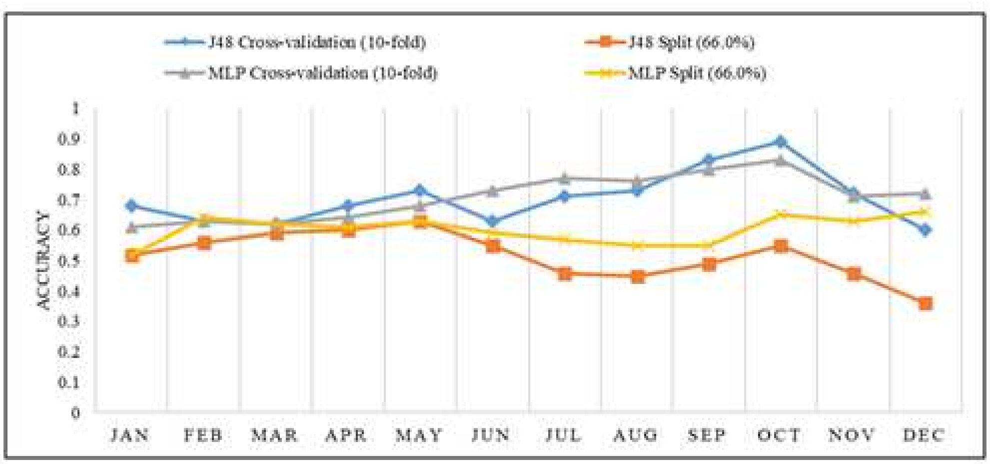


Figure 7

Month-wise Comparison of the Accuracy of the J48 (Cross-validation & Percent Split model) to the Multi-Layer Perceptron model for Accuracy for Sundargarh District

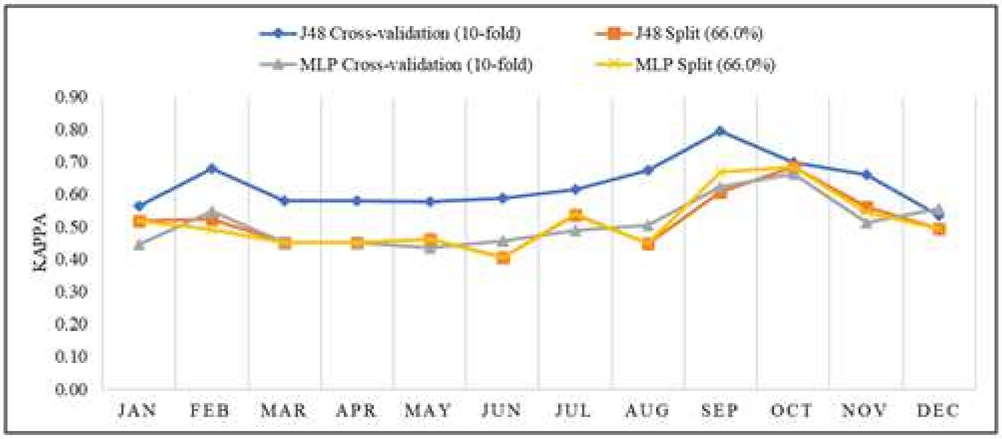


Figure 8

Month-wise Comparison of Kappa of the J48 (Cross-validation & Percent Split model) to the Multi-Layer Perceptron model for Accuracy for Sundargarh District

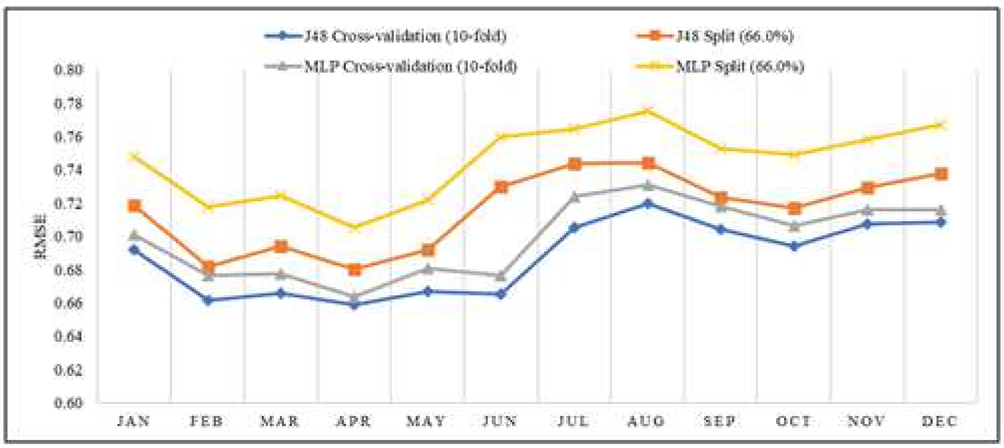


Figure 9

Month-wise Comparison of RMSE of the J48 (Cross-validation & Percent Split model) to the Multi-Layer Perceptron model for Accuracy for Sundargarh District

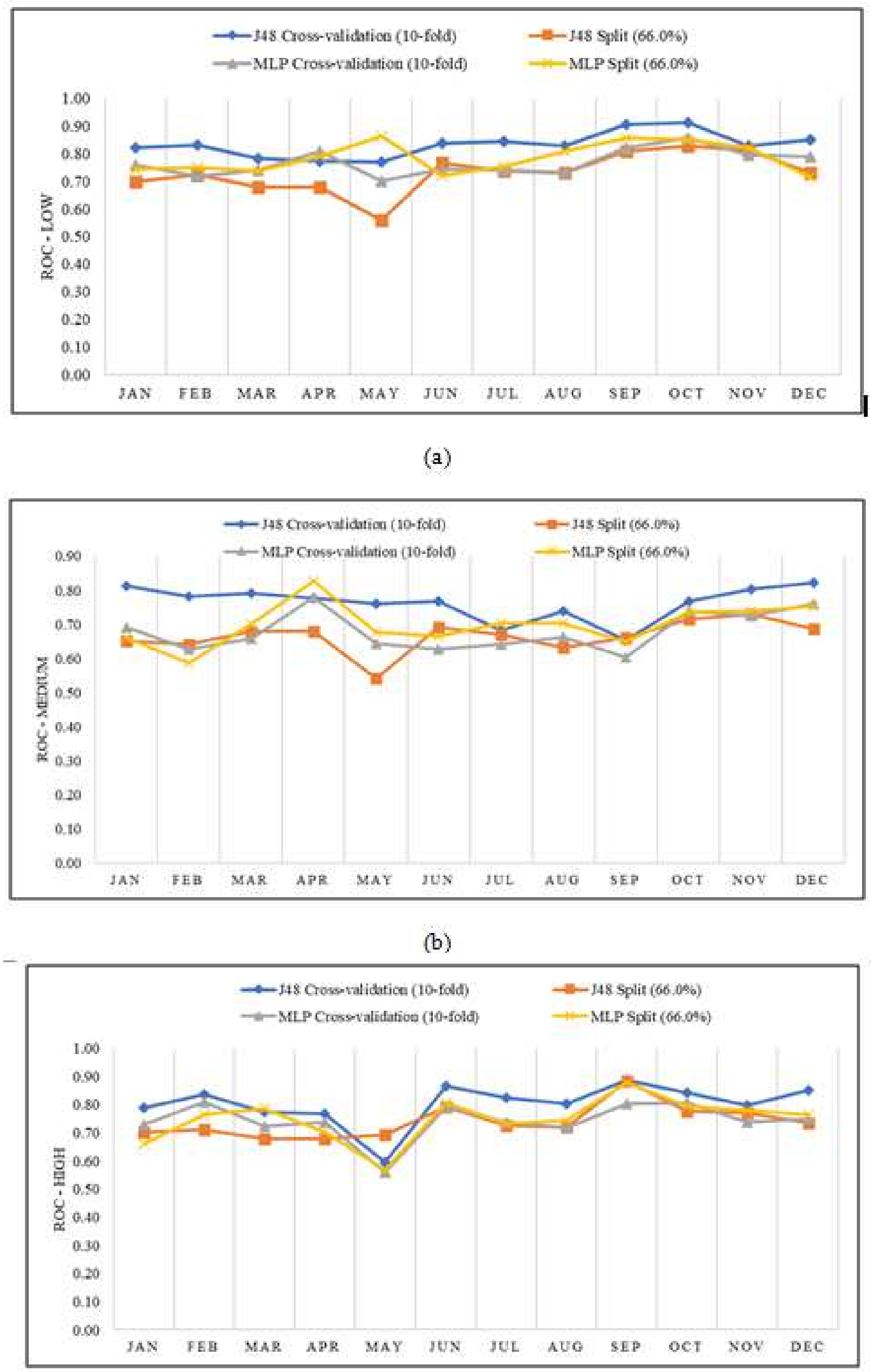


Figure 10

Month-wise Comparison of ROC for ((a) Low, (b) medium, and (c) High prediction category) of the J48 (Cross-validation & Percent Split model) to the Multi-Layer Perceptron model for Accuracy for Sundargarh District

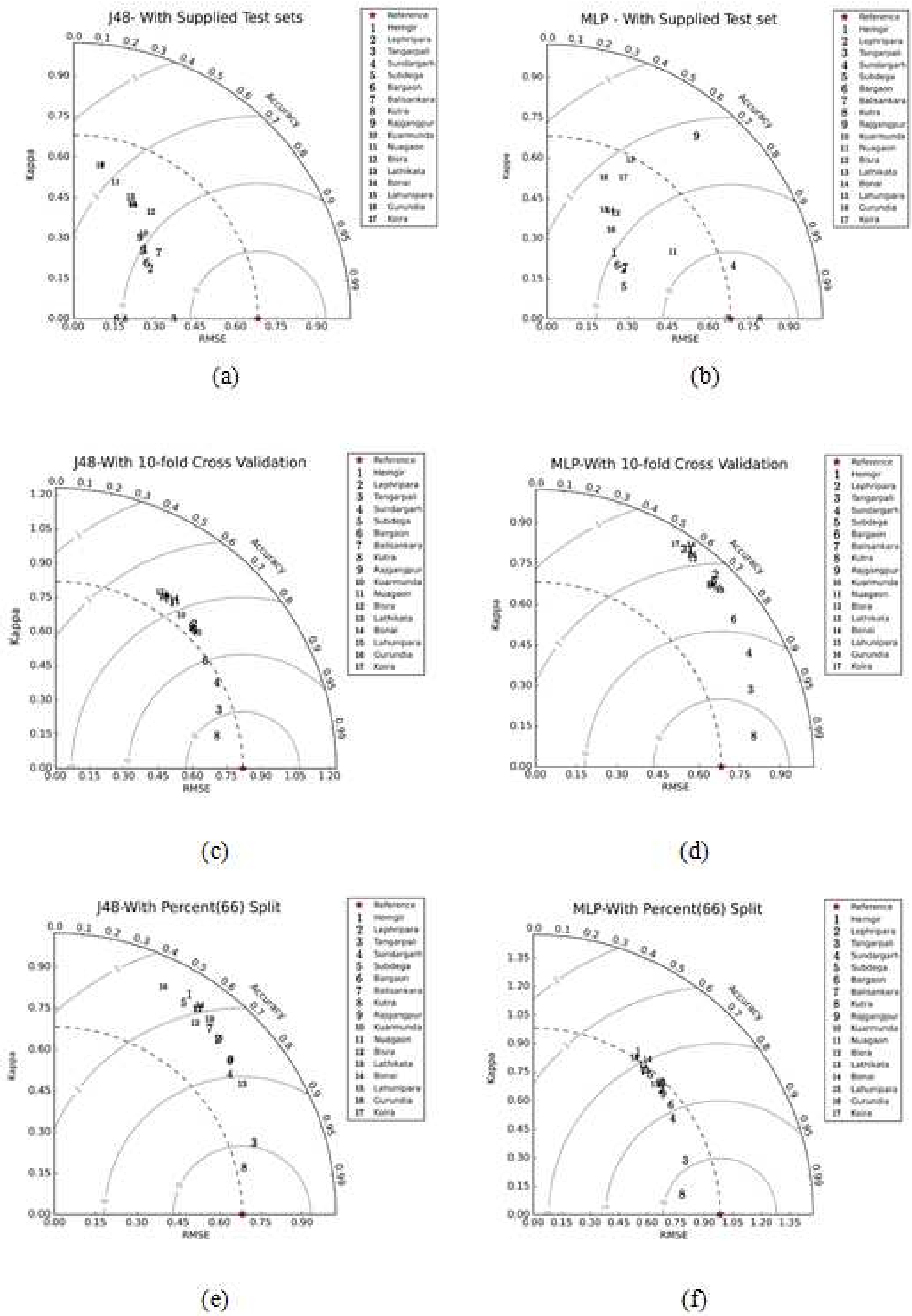
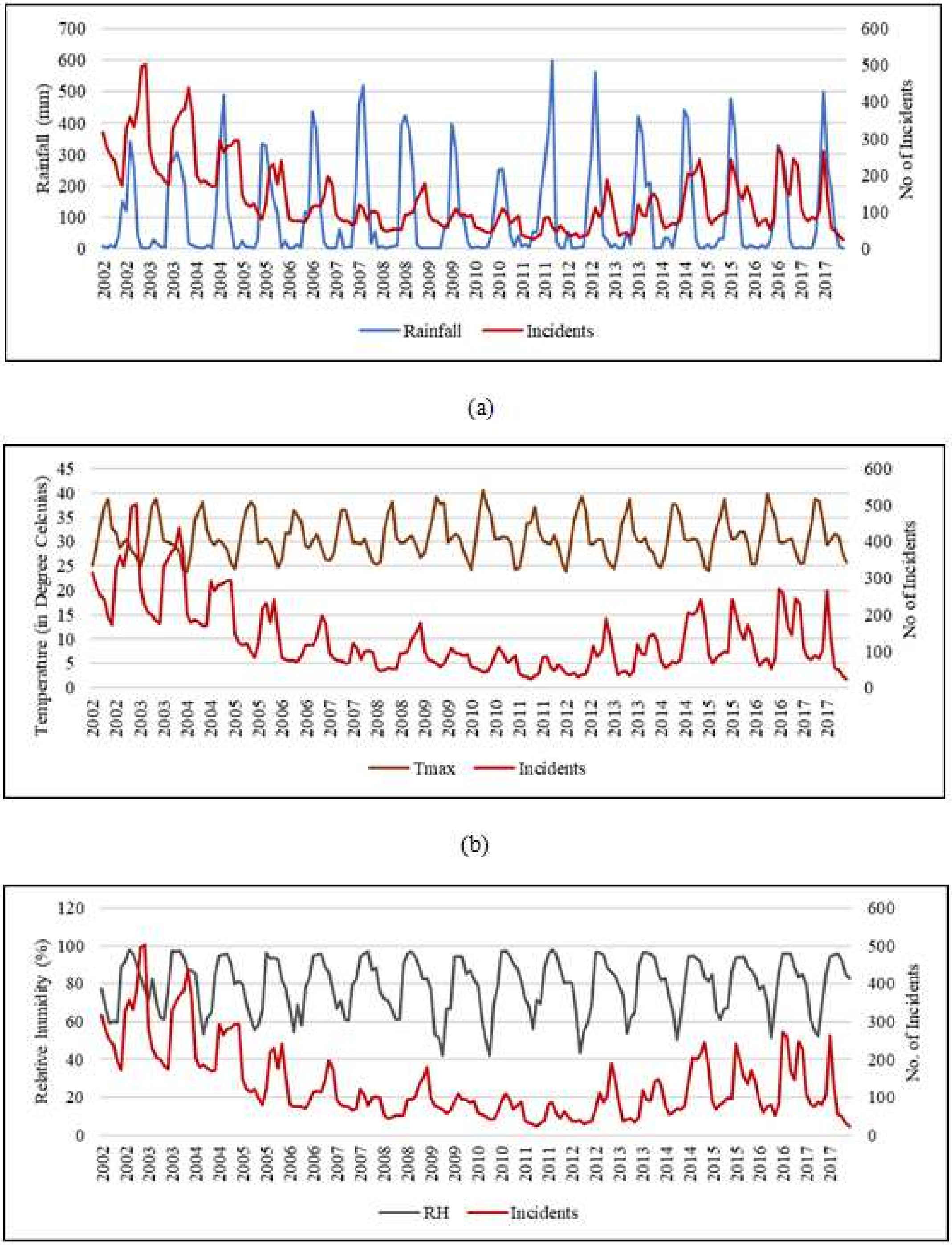


Figure 11

Block-wise accuracy (arc), RMSE (x-axis) and Kappa (Y-axis) for the J48 and the Multi-Layer Perceptron model with a Supplied test set, 10-fold Cross-validation & Percent Split classifiers



(c)

Figure 12

Time series plot for Monthly Rainfall (a), Maximum Temperature (b), and Relative Humidity (c) in comparison with malaria incidents for the period of 2002-2017