A MODEL FOR PREDICTING MALARIA OUTBREAK USING MACHINE LEARNING TECHNIQUE

Adebanji Stephen[1], Patrick O. Akomolafe[2], Kazeem I. Ogundoyin[1]

[1] Department of information and communication technology, University Osogbo, Osun State, Nigeria
[2] Department of computer science, University of Ibadan, Oyo State, Nigeria

Corresponding author: Adebanji Stephen, stephen.adebanji@uniosun.edu.ng

ABSTRACT: Malaria is a mosquito-borne infectious disease caused by protists (a form of microorganism) of the Plasmodium genus in humans and other animals. Malaria is a leading worldwide cause of morbidity and mortality. According to WHO the estimated value of malaria cases in 2019 was 229 million worldwide, with children under the age of 5 years having 67% (274,000) and being the most vulnerable group affected by malaria. Nigeria Demographics and health survey (NDHS) has a repository for data that can be used to predict malaria disease outbreak using machine learning techniques, from the literature reviewed no research has been carried out using machine learning technique to model the prediction of malaria outbreak using malaria incidence data from southwest Nigeria. This Research work used 5 supervised machine learning techniques to model the outbreak of malaria using meteorological and malaria incidence data of collected from 2010 - 2020, the machine learning techniques that was used are, Naive Bayes, Support Vector, Linear Regression, Logistic Regression, and K-Nearest Neighbor. The research was carried out using Scikit-learn Library that was imported into Anaconda IDE, the programming language used was Python programming language, The result of the research shows that Naive Bayes has the best accuracy for both testing and training with average accuracy of 79.1% and therefore is the best prediction model that can be used for predicting malaria incidence outbreak using the data set used in this research, Support Vector machine (SVM) is the second best prediction model that can be used for predicting malaria incidence outbreak for both testing and training data with average accuracy of 75.45%, followed by K-Nearest Neighbor with average accuracy of 70.8%, followed by Logistic Regression prediction model which has an average accuracy of 68%, based on this research work it is not advisable to use Linear Regression prediction model for predicting malaria incidence outbreak because it has an average accuracy of 26.05%.
KEYWORDS: Artificial Intelligence, Machine Learning, pre-processing, Prediction, Malaria, Support Vector machine, Naive Bayes, Logistic Regression

1. INTRODUCTION

Artificial intelligence (AI) is a field of computer science concerned with building smart machines capable of performing tasks that require human intelligence. one of the fields of artificial intelligence is machine learning, robotics, knowledge representation etc [7].

Machine learning is a field of artificial intelligence that trains algorithms to learn from previous experience and improve it to solve problems; machine learning has been a very important tool in solving problems that are related to Image Recognition, Speech Recognition, Medical diagnosis, and prediction of diseases. The prediction of disease outbreaks warns that a certain amount of disease may exceed the expected amount in the future at a particular time [5]. Prediction of infectious diseases attempts to predict the features of both seasonal epidemics and possible pandemics. Accurate and timely detection of infectious diseases by informing key preparedness and mitigation activities may help public health responses [3]. Malaria is a mosquito-borne infectious disease caused by protists (a form of microorganism) of the Plasmodium genus in humans and other animals. It starts with a bite from an infected female mosquito that introduces the protists into the circulatory system via its saliva, and eventually to the liver where they develop and reproduce. According to the WHO report of 2018, malaria is causing public health concern in developing countries, with estimated resultant deaths close to a million annually.

According to a publication by Severe Malaria Observatory on 28th of October 2020 Nigeria had the highest number of global malaria cases (25% of global malaria cases) in 2018 and accounted for the highest number of deaths, according to the 2019 World Malaria Report (24% of global malaria deaths). Malaria is transmitted throughout Nigeria; 76% of the population lives in areas of high transmission, while 24% of the population lives in areas of low transmission. In the south, the transmission season can last throughout the whole year and is about 3 months or less in the northern part of the country (Severe Malaria Observatory (SMO), 2020). Due to this high level of mortality rate, there is a need to develop an effective model that will predict malaria outbreak; the predictions by this model will assist health workers, government and hospitals to facilitate preventive measures earlier to the region with malaria outbreak thereby reducing the death rate caused by the outbreak of malaria [3].

2. LITERATURE REVIEW

2.1. Machine Learning

Machine learning is an application of Artificial intelligence that proves that systems have the ability to learn from previous experience and improve it without being programmed, and it is a technique that explores the analysis and creation of algorithms that can be used to make data predictions [4]. There are two categories of machine learning algorithms supervised and unsupervised learning algorithms.

2.2. Supervised Learning

Supervised learning algorithms are a machine learning algorithm that attempts to model relationships and dependencies between the output of the target prediction and the input features such that the output values for new data can be predicted based on those relationships that it has learned from previous data sets, supervised learning uses labeled data sets. Examples of supervised learning algorithms are Support Vector machine algorithm, K-Nearest Neighbor Algorithm, Naive Bayes Algorithm, Logistic Regression, Linear Regression e.t.c.

The algorithms used are Support Vector Machine (SVM), Naive Bayes algorithm, K-Nearest Neighbor algorithm (K-NN), Linear Regression algorithm (LiR), Logistic Regression algorithm (LoR) and Naive Bayes algorithm.

2.3. Support Vector Machine

SVM is a supervised machine learning algorithm which is usually used for classification or regression problems [9]. It has a unique technique that is called the kernel trick which it uses to transform our data and then based on these transformations it finds an optimal boundary between the possible outputs. Given the training data $(x_1, y_1), \dots, (x_n, y_n)$ where x_i is an element of X, the input value and y_i is an element of Y, the output value and n are the number of training data. The basic idea of SVM is to find.

$$f(x) = w.x + b$$

With at most e-deviation from the target value of y. Where w is the number of features represented in the training set, and w is the coefficient of x. This means that x and w are vectors while the statement above can be written mathematically as

$$f(x_i) - y < e$$

Where e represent a very small value

Also

$$f(x) = w_1 x_1 + w_2 x_2 + w_3 x_3 + \dots\dots + w_m x_m + b$$

The objective of the algorithm is to find the values of w and b such that the condition in the basic SVM equation is satisfied.

2.4 K- Nearest Neighbor

k Nearest Neighbor (or k-NN) is a supervised machine learning algorithm useful for prediction problems. It calculates the distance between the test data and the input and gives the prediction according [4].

How K-NN algorithm works

Step 1: It loads the data
Step 2: It then initialize K to our chosen number of neighbors for each example in the data
Step 3: After initializing the K to our chosen neighbors it then calculates the euclidean distance between first observation and new observation.
Step 4: And then add the distance and the index of the new observation to an ordered collection.
Step 5: After adding the distance and the index of the new observation it then sorts the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances.
Step 6: And pick the first K entries from the sorted collection.
Step 7: Then the algorithm will get the labels of the selected K entries.
Step 8: And return the mode of the K labels

2.5 Linear Regression

Linear regression is an attractive machine learning algorithm model because it has a very simple representation [11]. The representation of linear regression model is a linear equation that combines a specific set of input values (x) and the solution which is the predicted output for that set of output values (y). As such, both the input values (x) and the output value are numeric. The equations below show the prediction model equation for the linear regression model used in this project.

$$y = b_0 + b_1 x$$

Where y represents the prediction output, b0 represents the bias coefficient and b1 represents the coefficient for x and x is the input value for the model.

2.6 Logistic Regression

Logistic Regression is one of the most popular Machine Learning algorithms which are mostly used for classification problems; it is a predictive analysis algorithm and based on the concept of probability [7]. Logistic Regression is much similar to the Linear Regression, but Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. The Equation for logistic regression that is used in this project is show below

$$log[y/(1-y)] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \ldots b_nx_n$$

2.7 Naive Bayes

Naive Bayes is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

How Our Naive Bayes Algorithm works

- It loads the data and convert the data set into a frequency table
- It then creates a Likelihood table by finding the probabilities of the frequency table created.
- The algorithm will now use the Naive Bayesian equation shown below to calculate the posterior probability for each class. The class with the highest posterior probability is now the outcome of prediction.

Likelihood Class Prior Probability

$$P(c\,|\,x) = \frac{P(x\,|\,c)P(c)}{P(x)}$$

Posterior Probability Predictor Prior Probability

$$P(c\,|\,X) = P(x_1\,|\,c) \times P(x_2\,|\,c) \times \cdots \times P(x_n\,|\,c) \times P(c)$$

Where:

P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
P(c) is the prior probability of class.
P(x|c) is the likelihood which is the probability of a predictor given class.
P(x) is the prior probability of the predictor.

3. RELATED WORKS

[9] Developed a relationship between climate variables and a possible outbreak of malaria and also attempted to decide which algorithm is better suited to model the discovered relationship. For more than six years, they collected historical meteorological data and records of malarial cases and analyzed it using various classification techniques, such as KNN, Naive Bayes, and Extreme Gradient Boost. After evaluating the precision, the recall score, the accuracy score, the correlation coefficient for Matthews and the error rate for each case, they were able to find a few algorithms that perform well in malaria prediction and their research clearly show that, weather forecasts could be legitimately used to predict malaria outbreaks and probably take the steps required to avoid the loss of life due to malaria.

In a research by [8] they presented a clinical descriptive review of 165 patients from various age groups obtained from 2014to 2017 at Narasaraopet Medical Wards. They used the Synthetic Minority Oversampling Technique (SMOTE) to balance the class distribution, and then they conducted a comparative analysis on the dataset using the Naïve Bayesian algorithm on different platforms. 70 percent of the data was generated to train the Naive Bayesian algorithm from balanced class distribution data, and the rest of the data was used for testing the model for both weka and R programming environments. Their experimental findings showed that the weka environment classification of malaria disease data has the highest accuracy of 88.5% than the Naive Bayesian algorithm which has an accuracy of 87.5% using R programming language.

[6] used a model based on machine learning to identify the incidence of malaria using climate variability over a twenty-eight-year span across six Sub-Saharan African countries. The analysis starts with a process of feature engineering that determines the climate factors that influence the incidence of malaria, followed by the process of clustering k-means for outlier detection, and then the classification algorithm XGBoost. Their findings indicate that while the precise relationship between the occurrence of malaria and climate variability varies from one geographical area to another, non-seasonal changes in three climatic factors (precipitation, temperature, and surface radiation) contribute significantly to malaria outbreaks. Their system was compared with other classification models and the comparative findings revealed that other classification models outperformed their system.

[11] used decision tree classification algorithms on the WEKA workbench tool, a model was developed to predict the incidence of malaria in children between the ages of zero (0) and five (5) years. LMT, REPTree, Hoeffding tree, and J48 are the classification algorithms used. For the construction of the decision tree model, their research shows that the

J48 algorithm has greater output accuracy with the least margin of error.

One salient feature remained paramount in all the cases cited in the analysis of related works, and that is the fact that the ability of the different system to perform the expected task optimally , the research works show that there is a relationship between the meteorological data and malaria outbreak, and machine learning algorithms can be used to predict malaria outbreak using meteorological data and malaria incidence cases, with the researches reviewed different algorithms perform better in each cases for example in the reviewed work Naive Bayes perform better with an accuracy of 87.5% in another SVM perform better with an accuracy of 99% and in another research KNN perform better with an accuracy of 87%, but none of this research work has adopt the 5 algorithms we are using in this research work with our sourced data of malaria incidence cases from state region, to evaluate their performance in predicting malaria outbreak.

3.1 System Methodology

figure 1 shows the steps that are used in this research work, the research methodology metrological and malaria incidence data was collected from timezone.com and osun state ministry of health, the collected data was then cleaned and preprocessed, after preprocessing the data is divided into 70% training data set and 30% test data set, machine learning model (SVM, K-NN, LoR, LiR and Naive Bayes) is trained using the 70% training data and a prediction model was developed and tested using the remaining 30% of the preprocessed data, the result of the tested prediction model is evaluated using confusion matrix.

3.2 Data Acquisition

Meteorological data set obtained from www.timeanddate.com are the data used for modeling and analysis, comprising average temperature, average relative humidity, average wind speed on a monthly basis, also malaria incidence data set obtained from Osun State Ministry of Health Abere, Osun State, for all local governments in Osun State from 2010 to 2020. Support Vector Machine, Naive Bayes, K-NN, Linear Regression and Logistic Regression are the models used to analyze the collected data for malaria outbreak prediction.

3.3 Data Preprocessing

Data Preprocessing is the stage in which the data is translated or encoded to get it to such a state that it can now be easily parsed by the computer. In other words, the data's characteristics can now be readily interpreted by the algorithm. It involves Transformation and data cleaning, feature selection and data partitioning.

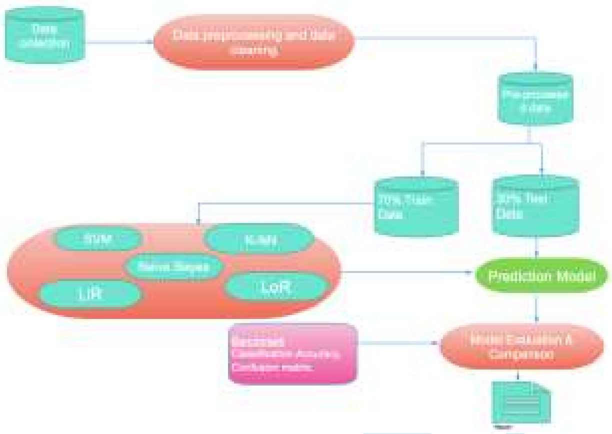

Figure 1 System Methodology

3.4 Data Transformation and data cleaning

Once the data has been obtained in the format of CSV files, the information relating to the years 2010 to 2020 was removed. We then summed the number of persons with confirmed uncomplicated Malaria, number of persons with severe malaria cases, and number of persons with clinically diagnosed malaria to get the number of malaria incidence per month, we calculated malaria incidence ratio by dividing the total number of malaria incidence per month by the estimated population for that year and we calculated the difference in the occurrence of the ratio using the equation (1)

$$diff = (maxratio - minratio)/30 \qquad eq(1)$$

$$ratio = tmc/p \qquad eq(2)$$

$$tmc = pcum + smc + pcdm \qquad eq(3)$$

where,
diff = difference in malaria incidence ratio.
tmc = total number of malaria cases
p = year population
ratio = ratio of malaria incidence
pcum = number of persons with confirmed uncomplicated malaria
smc = number of severe malaria cases seen
pcdm = number of persons with clinically diagnosed malaria

We then divided the risk level into 3

Low, when the ratio value is in the interval *minratio* and *minratio* + *diff,* for a given month.

Mild, when the ratio value is in the interval *minratio* + *diff* and *minratio* + 2 x *diff*, for a given month.

High, when the ratio value is in the interval *minratio* + 2 x *diff* and *maxratio* for a given month.

Finally, we added to the dataset the feature **"Outbreak"** as a dependent variable. Its values can be "Yes" if risk is "Mild" or "High" or "No" if risk has the value "Low".

3.5 Feature Selection

Feature selection process was done using correlation matrix as shown in figure 2, from correlation matrix shown in figure two the features used for training and testing the models are, Month, Average Temperature, Average relative humidity, Average precipitation, average wind speed, out patience attendance, number of malaria incidence and outbreak status. From the correlation matrix none of these features was strongly correlated hence they can all be used for the prediction model.

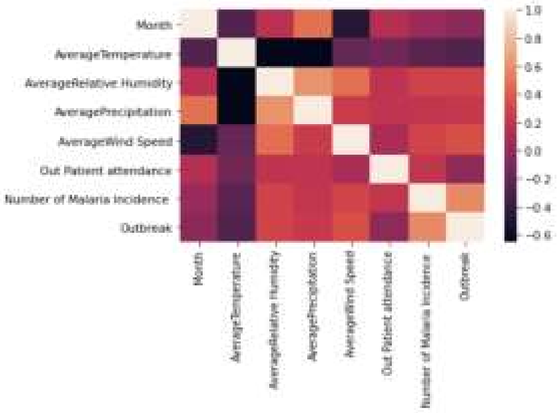

Figure 2. Image of data correlation matrix

3.6 Data Partitioning

The dataset was split into two parts: a sample containing 70% of the training data and 30% for the purpose of research. Then, using 5 main classification algorithms implemented in Python, the models were trained on the training sample: KNN, Support Vector Machine, Logistic regression, Linear Regression and Naive Bayes. On the 30% remaining data, the resulting models were tested, and the results were compared with the initial values of the Outbreak feature in the original dataset.

4. RESULTS

4.1 Result Presentation

The result of the different experiments carried out shows that logistic regression has a training accuracy of 81.9%, Linear Regression has a training accuracy of 45.6%, K-Nearest Neighbor (K-NN) has a training accuracy of 76.7%, Support Vector Machine (SVM) with linear Kernel has a training accuracy of 77.9% and Naive Bayes has a training accuracy of 82.6% During testing Logistic Regression has an Accuracy of 54.1%, K- Nearest Neighbor has an accuracy of 64.9%, Support Vector machine (SVM) has an accuracy of 73%, linear regression has an accuracy of 6.5% and Naive Bayes has accuracy of 75.6%, table 2 below.

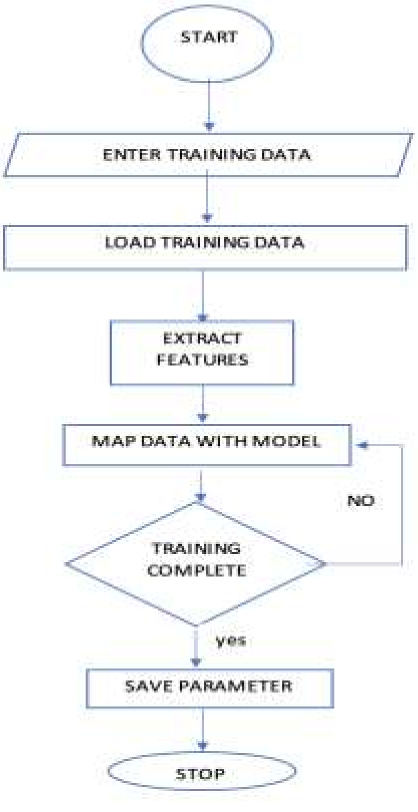

Figure 3. System flowchart for training

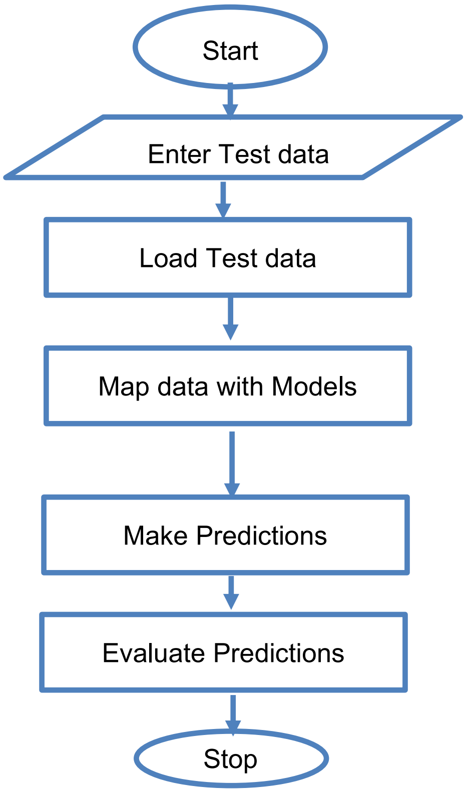

Figure 4 Flow chart for testing model

Table 2. Table of Accuracy comparison for the model algorithm

Prediction model	Training Accuracy	Test Accuracy	Average Accuracy
LoR	81.9	54.1	68
LiR	45.6	6.5	26.05
SVM	77.9	73	75.45
K-NN	76.7	64.9	70.8
Naive Bayes	82.6	75.6	79.1

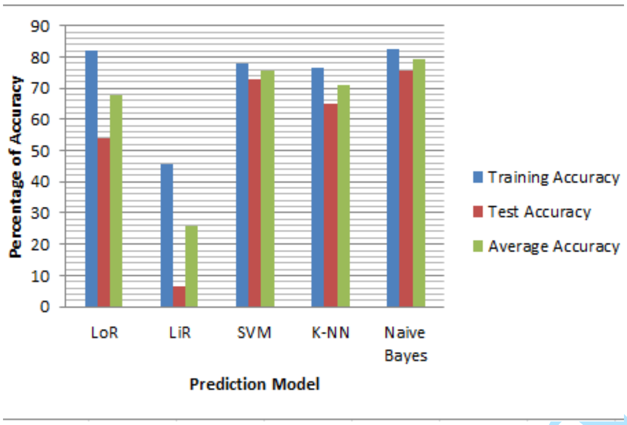

Figure 5 Image of Bar chart showing the comparison of prediction model algorithm

4.2 Discussion

From the result shown in the result representation section above, Naive Bayes has the best accuracy for both testing and training of 75.6% and 82.65 respectively and therefore is the best prediction model that can be used for predicting malaria incidence outbreak using the data set used in this research, Support Vector machine (SVM) can also be used because it is the second best prediction model for the data with 77.9% and 73% accuracy for training and testing respectively. The research work shows that it is not advisable to use Linear Regression prediction model for predicting malaria incidence outbreak using the same data used in this research work.

5. CONCLUSIONS

The research shows that Naive Bayes algorithm can be used to develop a model for predicting malaria outbreak using malaria incidence data and meteorological data. Experimental results were generated using Anaconda IDE with the Sk-learn Library for machine learning, The research also shows that linear regression algorithms should not be used to develop a model for predicting malaria outbreak. The performance of these algorithms were determined using a confusion matrix.

REFERENCES

[1]. **Adebayo Peter Idowu, Nneoma Okoronkwo and Rotimi E. Adagunodo** (2009), "Spatial Predictive Model for Malaria in Nigeria", Health Informatics in Developing Countries, Vol. 3 No. 2 pg 30 – 36.

[2]. **Ali Arab, Monica C. Jackson, and Cezar Kongoli** (2014), "Modelling the Effects of Weather and Climate on Malaria Distribution in West Africa", Malaria Journal, Vol. 13, pg. 126 – 135.

[3]. **Amuta E.U. and Houmsou R.S.** (2009), "Human Behavior and the Epidemiology of Parasitic Infections", African Journal of Pollution Health, Vol. 7(1), pg. 1 – 6.

[4]. **Babagana Modu, Nereida Polovina, Yang Lan, Savas Konur, Taufiq Asyhari A., and Yonghong Peng** (2017), "Towards a Predictive Analytics-Based Intelligent Malaria Outbreak Warning System" Applied Science, http://mpdi/journal/applsci, Vol. 7, pg. 836 – 856.

[5]. **Benjamin SC Uzochukwu, Ogochukwu P Ezeoke, Uloaku Emma-Ukaegbu, Obinna E Onwujekwe, and Florence T Sibeudu** (2010), "Malaria Treatment Services in Nigeria: A Review", Journal of the Nigeria Medical Association, Vol. 51, No. 3, pg 114 – 119.

[6]. **Godson Kalipe, Vikas Gauthum, and Rajat Kumar Behera** (2018), "Predicting Malaria Outbreak using Machine Learning and Deep Learning Approach; A Review and Analysis" Conference Paper, https://www.researchgate.net/publication/333492401

[7]. **Gurcan Comert, Negash Begashaw and Ayse Turhan-Comert** (2020), "Malaria Outbreak Dectection with Machine Learning Methods", https://doi.org/10.1101/2020.07.21.214213.

[8]. **Hamisu Ismail Ahmad** (2019), "Malaria Prediction using Bayesian and other Machine Learning Techniques", Unpublished, African Universities of Science and Technology, Department of Computer Science, Abuja, Nigeria.

[9]. **Marcin Cholewiński, Monika Derda, and Edward Hadaś** (2015), "Parasitic Diseases in Human Transmitted by Vectors", Annals of Parasitology, Vol. 61(3), pg. 137 – 157.

[10]. **Mengyang Wang, Hui Wang, Jiao Wang, Hongwei Liu, Rui Lu, Tongqing Duan, Xiaowen Gong, Siyuan Feng, Yuanyuan Liu, Zhuang Cui, Changping Li, and Jun Ma** (2019), "A novel model for malaria prediction based on ensemble algorithms", PLoS ONE 14(12): e0226910.
https://doi.org/10.1371/journal.pone.0226910.

[11]. **Odu Nkiruka, Rajesh Prasad, and Onime Clement** (2021), "Prediction of Malaria Incidence Using Climate Variability and Machine Learning", Informatics in Medicine Unlocked, Vol. 22. 100508.
https://www.sciencedirect.com/science/article/pii/S2352914820306596

[12]. **Olayinka T.C. and Chiemeke S.C.** (2019), "Predictive Pediatric Malaria Occurrence Using Classification Algorithm in Data Mining", Journal of Advances in Mathematics and Computer Science, 31(4) No. 39029, pg. 1 – 10.
https://doi.org/10.9734/jamcs/2019/v31i430118

[13]. **Opeyemi A. Abisoye and Rasheed G. Jimoh** (2018), "Comparative Study on the Prediction of Symptomatic and Climate based Malaria Parasite Counts using Machine Learning Models", I.J. Modern Education and Computer Science, Vol. 4, pg 18.25.
http://j.mecs-press.net/ijmecs/ijmecs-v10-n4/IJMECS-V10-N4-3.pdf

[14]. **Rachel N. Bronzan, Meredith L. McMorrow and S. Patrick Kachur** (2008), "Diagnosis of Malaria", Mol Diag Ther, Vol. 12 (5), pg. 299 – 306.
https://doi.org/10.1007/BF03256295

[15]. **Sajana T. and Narasingarao M.R.** (2018), "Classification of Imbalanced Malaria Disease Using Naïve Bayesian Algorithm", International Journal of Engineering & Technology, Vol. 7(2.7), pg 786 – 790.

[16]. **Samy S. Abu Naser and Suheir H. ALmursheidi** (2016), "A Knowledge Based System for Neck Pain Diagnosis", Worldwide Journal of Multidisciplinary Research and Development, Vol. 2(4), pg. 12 – 18.

[17]. **Second Rural Access and Mobility Project (RAMP-2).** (2015), Abbreviated Resettlement Action Plan; Construction /Rehabilitation of Prioritized Rural Roads and Rivers Crossing, Federal Ministry of Agriculture and Rural Development, Osun State. SFG1467 V3.

[18]. **Srivastava N.** (2005), "A logistic Regression Model for Predicting the Occurrence of Intense Geomagnetic Storms", Annales Geophysicae, Vol. 23, pg 2969 – 2974.

[19]. **Viktor Andersson** (2017), "Machine Learning in Logistics: Machine Learning", Unpublished, Lulea University of Technology, Department of Computer Science, Electrical and Space Engineering.

[20]. **United States Embassy in Nigeria** (2011), "Nigeria Malaria Fact Sheet", Economic Section, United States Embassy in Nigeria. http://nigeria.usembassy.gov

[21]. **Vijeta Sharma, Ajai Kumar, Lakshmi Panat, Dr. Ganesh Karajkhede, and Anuradha Lele** (2016), "Malaria Outbreak Prediction Model Using Machine Learning", International Journal of Advanced Research in Computer Engineering & Technology, Vol. 4 Issue. 12, pg 4415 – 4419.