CARITAS UNIVERSITY AMORJI-NIKE, EMENE, ENUGU STATE

Caritas Journal of Physical and Life Sciences

CJPLS, Volume 3, Issue 1 (2024)

Article History: Received: 12th March, 2024 Revised: 17th april, 2024 Accepted: 7th May, 2024

The Potential of Machine Learning Techniques in Predicting Malaria Outbreaks in Nigeria

[1]Okoronkwo, M.C.,
Omankwu, Obinnaya.C.B.,
Kanu, Chigbundu
Department of Computer Science, Michael Okpara University of Agriculture, Umudike
saintbeloved@yahoo.com

Abstract

Malaria, an infectious disease transmitted by mosquitoes and caused by protists of the Plasmodium genus, poses a significant global health threat, contributing substantially to morbidity and mortality rates. World Health Organization (WHO) estimated approximately 229 million cases worldwide, with children under five years old comprising 67% (274,000) of those affected, representing the most vulnerable demographic group. Despite the prevalence of malaria, existing research has not extensively explored the utilization of machine learning techniques to predict malaria outbreaks, in Nigeria. This study aims to fill this gap by employing five supervised machine learning methods: Naive Bayes, Support Vector Machines (SVM), Linear Regression, Logistic Regression, and K-Nearest Neighbor. Utilizing meteorological and malaria incidence data spanning from 2010 to 2020, the research employed the Scikit-learn library within the Anaconda IDE, utilizing the Python programming language. Results indicate that Naive Bayes achieved the highest accuracy, with an average accuracy of 79.1% for both testing and training datasets, making it the optimal model for predicting malaria incidence outbreaks based on the dataset utilized. Following closely is Support Vector Machine (SVM) with an average accuracy of 75.45% for both testing and training data, followed by K-Nearest Neighbor with an average accuracy of 70.8%. Logistic Regression exhibited an average accuracy of 68%. However, Linear Regression, with an average accuracy of 26.05%, is not recommended for predicting malaria incidence outbreaks based on the findings of this research.

Keywords: ***Artificial Intelligence, Machine Learning, Malaria, Naive Bayes, Prediction, Support Vector machine***

1. INTRODUCTION

Artificial intelligence (AI) encompasses a branch of computer science dedicated to constructing intelligent machines capable of mimicking human cognitive functions. Among its various domains are machine learning, robotics, and knowledge representation [7]. Machine learning, a subset of AI, involves training algorithms to glean insights from past experiences and refine their performance to tackle complex problems. It has emerged as a pivotal tool in addressing challenges ranging from image and speech recognition to medical diagnosis and disease prediction.

The prediction of disease outbreaks is a critical application of machine learning, offering insights into the likelihood of disease surpassing anticipated levels at specific times [5]. This predictive capacity is particularly crucial in the realm of infectious diseases, where timely detection can inform preparedness and mitigation efforts, thereby bolstering public health responses [3]. Malaria, a mosquito-borne illness caused by Plasmodium protists, poses a significant global health threat, with transmission occurring via infected mosquito bites, leading to approximately a million deaths annually, predominantly in developing countries [8].

Nigeria, in particular, bears a heavy burden of malaria, comprising 25% of global malaria cases and deaths in 2018 [8]. With 76% of its population residing in high-transmission areas, the country experiences varying transmission seasons across regions [9]. Given the profound impact of malaria on mortality rates, there exists a pressing need to develop effective predictive models to anticipate outbreaks. Such models can empower healthcare professionals, governmental bodies, and hospitals to implement preemptive measures in regions prone to malaria outbreaks, thereby mitigating the associated mortality rates [3].

2. LITERATURE REVIEW

2.1. Machine Learning

Machine learning, a subset of Artificial Intelligence, demonstrates the capability of systems to autonomously learn from past experiences and enhance their performance without explicit programming. It involves the development and analysis of algorithms designed to make predictions based on data [4]. Machine learning algorithms can generally be categorized into two types: supervised and unsupervised learning algorithms.

2.2. Supervised Learning

Supervised learning algorithms are a subset of machine learning techniques designed to discern patterns and correlations between input features and the target output. By leveraging labeled datasets, these algorithms aim to construct models that can predict the output values for new data based on the learned relationships from previous data instances. Examples of supervised learning algorithms include Support Vector Machines, K-Nearest Neighbors, Naive Bayes, Logistic Regression, and Linear Regression, among others. The utilized algorithms comprise Support Vector Machine (SVM), Naive Bayes, K-Nearest Neighbor (K-NN), Linear Regression (LiR), and Logistic Regression (LoR).

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 ($_1$, $_1$), . . . , (,) where is an element of X, the input value and is an element of Y, the output value and n are the number of training data. The basic idea of SVM is to find.

$$(\ ) = \ . \qquad +$$

With at most -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 Where represent a very small value

Also

$(x) = w1x1 + w2x2 + w3x3+. \ldots \ldots +wmxm + b$

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

2.4 K- Nearest Neighbor

The k Nearest Neighbor algorithm (k-NN) is a supervised machine learning technique utilized for prediction tasks. It involves computing the distance between the test data and the input, then making predictions accordingly [4].

Here's how the k-NN algorithm operates:

1. Data Loading: The algorithm first loads the dataset.
2. Initialization: It initializes K to the chosen number of neighbors for each example in the dataset.
3. Distance Calculation: After setting K, the algorithm computes the Euclidean distance between the first observation and the new observation.
4. Collection Addition: It adds the distance and the index of the new observation to an ordered collection.
5. Sorting: Subsequently, the algorithm sorts the ordered collection of distances and indices from smallest to largest (in ascending order) based on the distances.
6. Selection: The algorithm selects the first K entries from the sorted collection.
7. Label Retrieval: It retrieves the labels of the selected K entries.
8. Prediction: Finally, the algorithm returns the mode of the K labels.

2.5 Linear Regression

Linear regression stands out as an appealing machine learning algorithm due to its straightforward representation [11]. This representation takes the form of a linear equation, combining a designated set of input values (x) with the predicted output (y) for that specific set of inputs. Consequently, both the input (x) and output (y) values are numeric. The equations below depict the prediction model equation for the linear regression model employed in this project.

The equation for the linear regression model used in this project is:

$y = mx + b$

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

2.6 Logistic Regression

Logistic Regression ranks among the most widely utilized Machine Learning algorithms, primarily employed for classification tasks; it operates on the principles of probability [7]. While akin to Linear Regression, which addresses regression problems, Logistic Regression focuses on solving classification challenges. It forecasts the outcome of a categorical dependent variable, necessitating discrete values such as Yes or No, 0 or 1, true or false, etc. Rather than delivering exact values of 0 and 1, Logistic Regression provides probabilistic values ranging between 0 and 1. The equation utilized for logistic regression in this project is displayed below:

$\log[y/(1 - y)] = b0 + b1x1 + b2x2 + b3x3... bnxn$

Naive Bayes:

Naive Bayes functions as a classification method grounded in Bayes' Theorem, underpinned by the assumption of independence among predictors. Essentially, a Naive Bayes classifier presumes that the presence of a specific feature within a class bears no relation to the presence of any other feature.

Operation of Our Naive Bayes Algorithm: • Data loading and conversion into a frequency table are initiated. • A Likelihood table is constructed by determining the probabilities derived from the frequency table. • Subsequently, the algorithm employs the Naive Bayesian equation below to compute the posterior probability for each class. The class exhibiting the highest posterior probability emerges as the prediction outcome.

The Naive Bayes equation is:

$P(Y|X) = P(X|Y) * P(Y) / P(X)$

Where:

- P(Y|X) is the posterior probability of class Y given predictor variables X.
- P(X|Y) is the likelihood of predictor variables X given class Y.
- P(Y) is the prior probability of class Y.
- P(X) is the prior probability of predictor variables X.

3. EMPIRICAL REVIEW

[9] Explored the correlation between climate variables and potential malaria outbreaks while assessing the efficacy of various algorithms for modeling this relationship. Over a span of six years, they amassed historical meteorological data and malaria case records, employing multiple classification techniques including K-Nearest Neighbors, Naive Bayes, and Extreme Gradient Boost. Through evaluation metrics such as precision, recall score, accuracy score, Matthews correlation coefficient, and error rate, they identified algorithms conducive to accurate malaria prediction. Their findings underscore the viability of weather forecasts in predicting malaria outbreaks, thereby aiding in preventive measures to mitigate malaria-related fatalities.

In another study by [8], a clinical descriptive review involving 165 patients across diverse age groups from 2014 to 2017 at Narasaraopet Medical Wards was presented. Utilizing the Synthetic Minority Oversampling Technique (SMOTE) to rectify class imbalances, the researchers conducted a comparative analysis employing the Naïve Bayesian algorithm across different platforms. Seventy percent of the data was allocated for training the algorithm with balanced class distribution, while the remainder was utilized for testing in both the Weka and R programming environments. Experimental results revealed that the Weka environment yielded the highest accuracy of 88.5% in classifying malaria disease data, surpassing the Naive Bayesian algorithm's accuracy of 87.5% using the R programming language.

[6] Leveraged machine learning to discern malaria incidence patterns vis-à-vis climate variability spanning twenty-eight years across six Sub-Saharan African countries. Employing feature engineering to identify climate factors influencing malaria incidence, followed by outlier detection using k-means clustering and classification utilizing XGBoost algorithm, their investigation elucidated the substantial impact of non-seasonal variations in

precipitation, temperature, and surface radiation on malaria outbreaks. Comparative analysis with other classification models demonstrated superior performance of alternative models over theirs.

DECISION METHODOLOGY

Meteorological and malaria incidence data were procured from timezone.com and the Osun State Ministry of Health, respectively. Following data acquisition, cleaning, and preprocessing, the dataset was partitioned into a 70% training set and a 30% test set. Machine learning models including Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), Logistic Regression (LoR), Linear Regression (LiR), and Naive Bayes were trained on the 70% training data, and a predictive model was developed and tested using the remaining 30% of preprocessed data. Model performance was evaluated using confusion matrix analysis.

Data Acquisition

Meteorological data sourced from www.timeanddate.com encompassed average temperature, relative humidity, and wind speed on a monthly basis, while malaria incidence data obtained from the Federal Ministry of Health spanned across all the State from 2010 to 2020. These datasets served as inputs for modeling and analysis, with the aforementioned machine learning models applied to predict malaria outbreaks.

Data Processing

Data preprocessing entailed transformation, cleaning, feature selection, and partitioning. Transformation and cleaning operations involved converting CSV files to remove data for years outside the 2010-2020 range and calculating malaria incidence ratios based on confirmed uncomplicated malaria cases, severe malaria cases, and clinically diagnosed malaria cases per month. Feature selection utilized correlation matrix analysis to identify relevant predictors for training and testing the models, including month, average temperature, relative humidity, precipitation, wind speed, outpatient attendance, number of malaria cases, and outbreak status.

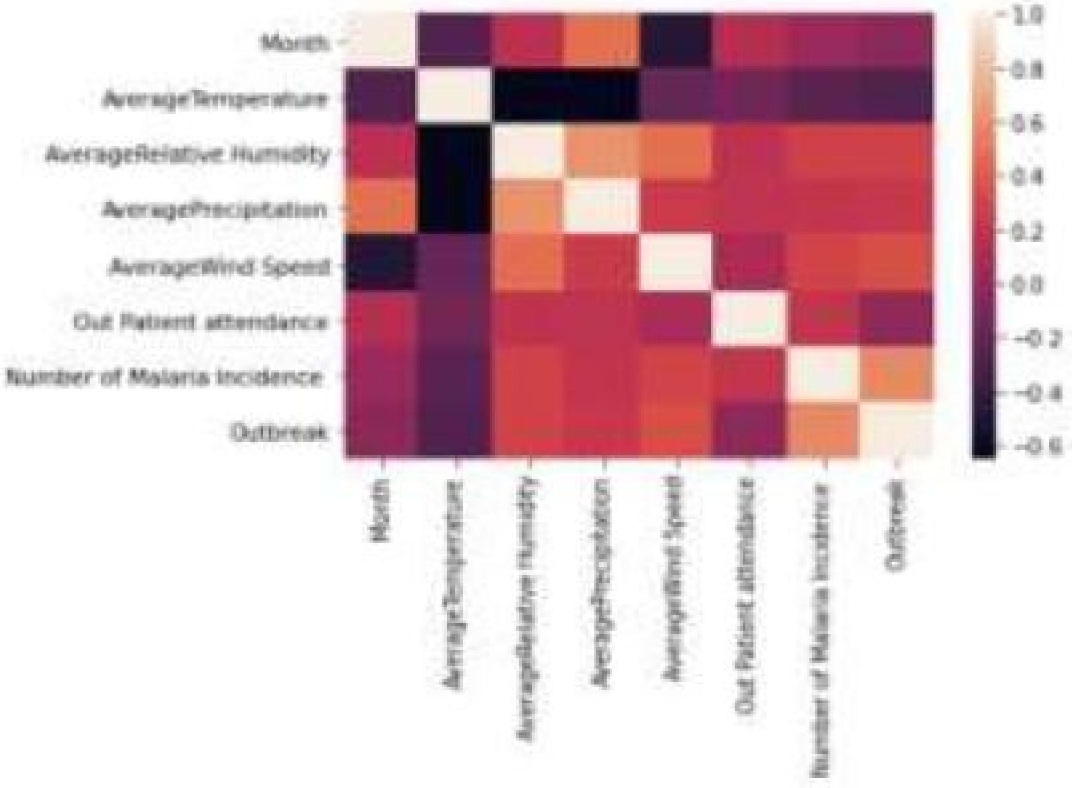


Figure 2. Image of data correlation matrix

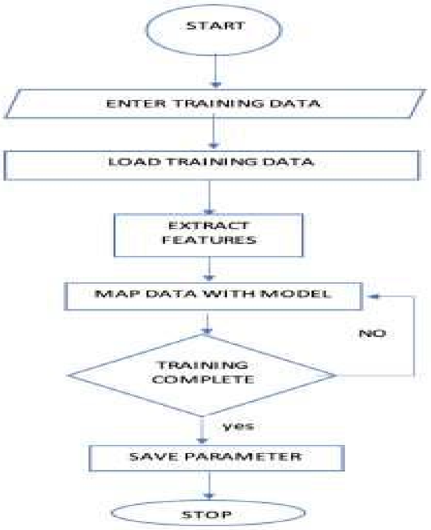


Figure 3. System flowchart for training

3.6 Data Partitioning

The dataset underwent partitioning into two segments: a training sample comprising 70% of the data and a 30% subset reserved for research purposes. Subsequently, employing five primary classification algorithms implemented in Python—K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Logistic Regression, Linear Regression, and Naive Bayes—the models were trained on the training sample. Following this, the resulting models were tested on the remaining 30% of the data, and the outcomes were compared with the initial values of the Outbreak feature in the original dataset.

RESULTS

Presentation of Results

The results obtained from the various experiments conducted reveal that Logistic Regression exhibited a training accuracy of 81.9%, Linear Regression achieved a training accuracy of 45.6%, K-Nearest Neighbor (KNN) attained a training accuracy of 76.7%, Support Vector Machine (SVM) with a linear kernel yielded a training accuracy of 77.9%, and Naive Bayes demonstrated a training accuracy of 82.6%. During testing, Logistic Regression recorded an accuracy of 54.1%, K-Nearest Neighbor achieved an accuracy of 64.9%, Support Vector Machine (SVM) scored 73% accuracy, Linear Regression yielded an accuracy of 6.5%, and Naive Bayes exhibited an accuracy of 75.6%, as depicted in Table 2 below.

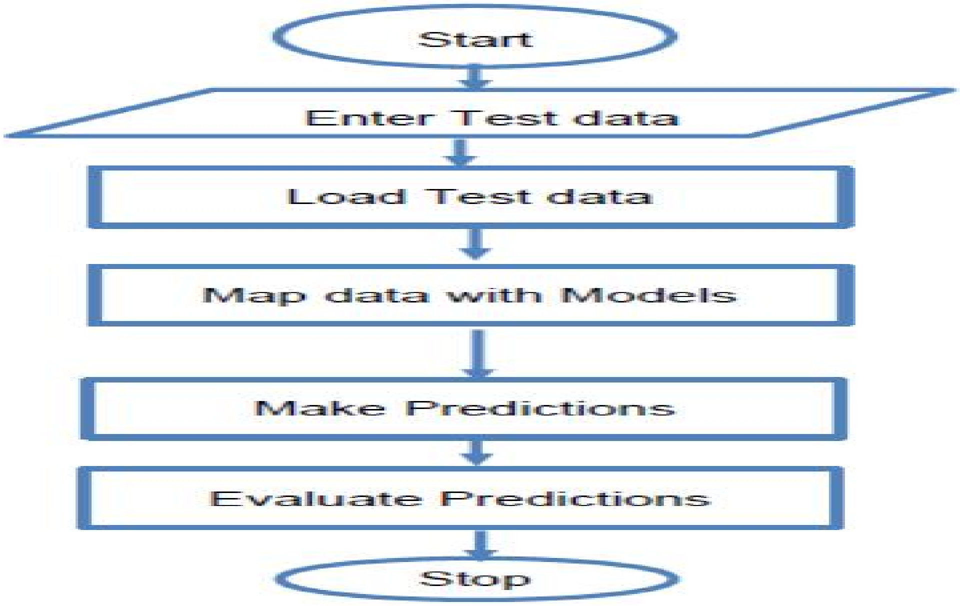
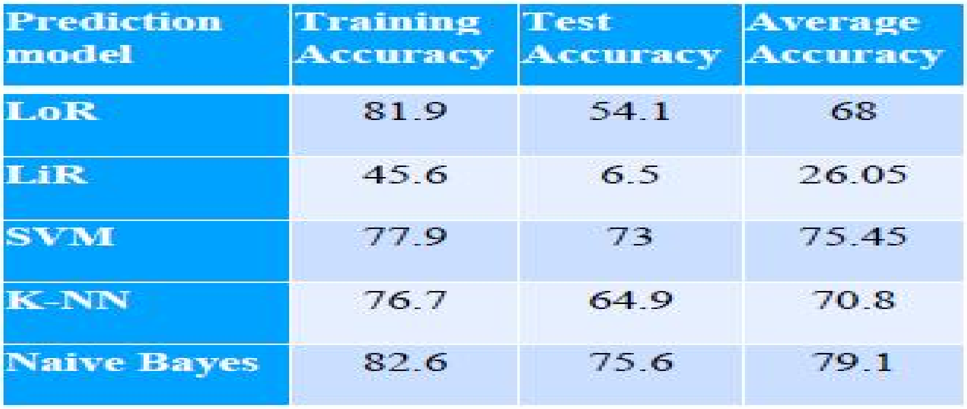


Figure 4 Flow chart for testing model

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

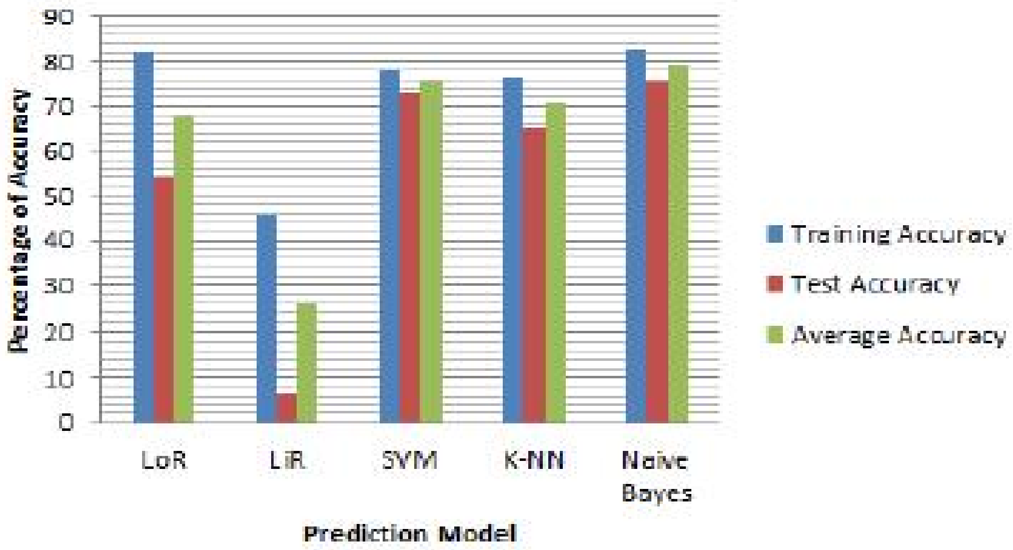


Figure 5 Image of Bar chart showing the comparison of prediction model algorithm

4.2 Discussion

As indicated in the results presented above, Naive Bayes exhibits the highest accuracy for both training and testing, achieving 82.65% and 75.6% respectively. Consequently, it emerges as the optimal prediction model for forecasting malaria incidence outbreaks using the dataset employed in this study. Additionally, Support Vector Machine (SVM) emerges as a viable option, ranking as the second-best prediction model with accuracies of 77.9% and 73% for training and testing respectively. However, it is noteworthy that utilizing Linear Regression for predicting malaria incidence outbreaks with the same dataset is not recommended based on the findings of this research.

CONCLUSIONS

The research underscores the efficacy of the Naive Bayes algorithm in developing predictive models for malaria outbreak prediction utilizing both malaria incidence and meteorological data. Experimental results were obtained using Anaconda IDE with the Sk-learn Library for machine learning. Furthermore, the study advises against employing linear regression algorithms for this purpose, with the performance of the algorithms assessed 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.