A Symptom-Based Machine Learning Model for Malaria Diagnosis in Nigeria

2021 9th International Symposium on Digital Forensics and Security (ISDFS) | 978-1-6654-4481-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISDFS52919.2021.9486315

Bilyaminu Muhammad
Department of Computer Science
Usmanu Danfodiyo University, P.M.B, 2346, Sokoto, Nigeria
bilyaminu49@gmail.com
ORCID: 0000-0003-4281-5729

Asaf Varol
Department of Computer Engineering
College of Engineering and Natural Sciences
Maltepe University,34857 Maltepe/Istanbul, Turkey
asafvarol@maltepe.edu.tr
ORCID: 0000-0003-1606-4079

***Abstract*— Malaria, with around 200 million cases worldwide, tends to kill more people than war and crises. With efforts to reduce mortality rates being futile, an inadequate malaria diagnosis is one of the barriers to a successful reduction in mortality. Machine learning methods were used to classify the stages of malaria in patients to improve diagnosis. To predict the stages of malaria, this research used knowledge of an algorithm of machine learning for a predictive model. A 77% accurate decision algorithm was developed using the symptoms of patients to identify their malaria stages. This research also discovered that malaria does not kill only children (between 0–5 years), in contrast to what has been pointed out in many research studies. This study shows that older women are more likely to experience severe stages of malaria. Therefore, adequate care should be considered for these women once they show some of the significant symptoms as described in the model. This approach applies to everyone with the symptoms set out in the model. This system will provide a preliminary test before conducting a confirmatory diagnosis in the laboratories.**

Keywords*— *Malaria, Sokoto, Decision tree, Machine Learning

I. INTRODUCTION

Malaria is caused by infected female Anopheles mosquito bites that infect red blood cells with the genus Plasmodium transmitted by protozoan parasites. The causes of malaria in humans are five Plasmodium species: Plasmodium falciparum, Plasmodium vivax, Plasmodium malariae, Plasmodium ovale, and Plasmodium knowlesi. P. falciparum and P. vivax are the two most common species. P. falciparum is the most severe form of malaria and is responsible for the majority of deaths worldwide [1].

P. falciparum is the most widespread malaria parasite in sub-Saharan Africa, accounting for 99% of estimated cases of malaria in 2016. Outside of Africa, P. vivax is the leading parasite, constituting 64% of malaria cases in America, over 30% of malaria cases in Southeast Asia, and 40% east of the Mediterranean [2]

In non-severe malaria, young stages of P. falciparum (24-hours-old) are often present in the peripheral blood, while in the peripheral blood, all stages of severe malaria may be present. To aid in solving this problem, the algorithm is simple to interpret, using a sample of patients diagnosed with different phases of malaria in one of the renowned hospitals. The algorithm will be useful in creating a model for discovering patient malaria stages. It will also assist in decision making.

Most research has shown that children are prone to dying from malaria, particularly in Africa, where a child dies from malaria almost every minute. The 2016 World Malaria Report [1] shows that an estimated 3.2 billion people are at risk of malaria infection and of developing illness in 95 countries and territories, and 1.2 billion are at high risk (in a year, 1 out 1000 have a chance of becoming infected with malaria). Around 214 million malaria cases occurred worldwide in 2016, and some 438,000 malaria deaths were reported. The casualties in Africa were higher and were estimated at 92% [3] of all malaria deaths; children under the age of five accounted for more than two-thirds of all deaths. Common malaria symptoms include fever, tiredness, headaches, and, in some severe cases, seizures and comas, which can lead to death.

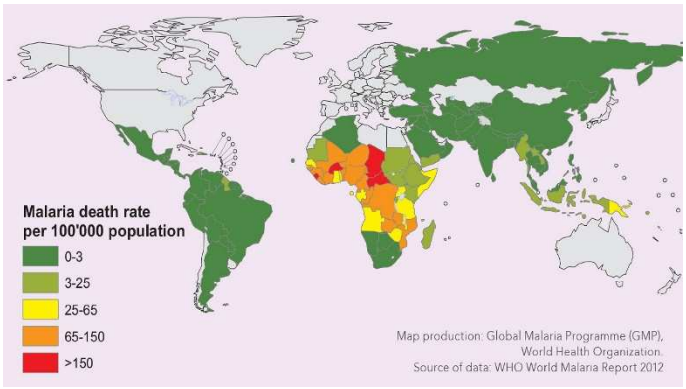

Fig. 1. Worldwide malaria death rates (Source: WHO World Malaria Report, 2012)

Cases of incorrect diagnosis lead to erroneous diagnostic decisions. With false-negatives, antibiotics are unnecessary, the second consultation, working hours have been lost, and in certain cases, severe malaria progresses. With false-positives, an incorrect diagnosis leads to the use of anti-malarial drugs, and there may there be side effects such as fatigue, stomach pain, diarrhea, vomiting, and severe complications.

This careful examination of a malaria diagnosis has contributed to efforts to diagnose malaria automatically.

This paper focuses more on the implementation of a symptom-based machine learning model for malaria diagnosis. The rest of this paper comprises of relevant literature review in section II. Section III discusses the Materials and Methods used for the research. Detail of the proposed model in section IV. Finally, the conclusion of the paper in section V.

II. Related Works

The approaches to predictions include statistical modeling, mathematical modeling, and machine learning [4]. In terms of predictions and decision making, prediction models play an essential role.

In the recent past, machine learning (ML) in the health sciences [5] has been used to diagnose various diseases, including cancer [6].

ML finds the correct formulation and safe medications for disabling a disease virus in pharmacology [7]. Effective treatment is also used to select ML. In agriculture, as with plant predictions, ML can also be used to increase production in agriculture [8]. ML is used in the corporate world as a prediction for the movement of stock and stock price index [9].

Malaria is currently predicted in many nations using environmental risk factor data (e.g., climate conditions) to forecast the incidence of malaria over a certain period in a geographical area [10].

An automated malaria parasite diagnosis was proposed in 2015 with the neural network and support vector machine. Errors in manual diagnostics and time consumption are unavoidable. An artificial neural network (ANN) classifier provides 80% accuracy for the affected and 77%for the unaffected, and a support vector machine (SVM) provides 90% accuracy for the affected and 100% accuracy for the unaffected. But the researchers did not know that a classifier's performance depends on the domain being considered. The study concentrates on processing asymptomatically. The impacts of symptomatic and climatic conditions were not considered [11].

In 2015, the malaria outbreak prediction model using machine learning was proposed. Early prediction of an outbreak of malaria is the key to malaria morbidity control and can assist different health organizations better to target medical resources to the areas of greatest need. For malaria prediction, two popular data mining classification algorithms were used: SVM and ANN. The parameters used in the binary values “Yes” or “No” are average monthly rainfall, temperature, humidity, the total number of positive cases, the total number of cases of P. falciparum, and the outbreak. The SVM model can predict the outbreak in advance for 15–20 days. The prediction accuracy needs to be improved by using more training data. Also, in the model, the individual positive cases need to be considered as one of the training and testing features, not the total number of positive cases [12].

Applying different predictive methods to the same data and exploring the predictive capacity of environmental and non-environmental variables, including intervention reduction transmission and the use of common predictive accuracy measures, will enable malaria researchers to compare and improve models and methods that should improve the quality of malaria prediction [13].

An analytical program was initiated to detect individual malaria parasite count from a complex network of several infection counts. The experiment was conducted with around 1,200 data points, and two classifiers, SVM and ANN, were used for prediction. Experimental results showed that SVM produced the following results: accuracy 85.60%, sensitivity 84.06%, specificity 86.49%, false positive rate (FPr) 0.1351%, and false-negative rate (FNr) 0.1594%. The neural network model, in comparison, had the following results: accuracy 48.33%, sensitivity 60.61%, specificity 45.48%, false positive rate (FPr) 0.5442%, and false-negative rate (FNr) [14].

III. Materials And Methods

A total of 500 samples from hospitals were collected together with the symptomatic characteristics of experimental laboratory data from patients where 300 samples are used for training, 150 samples for testing and 50 samples for validation. At the Maryam Abacha Hospital Sokoto, Nigeria, demographic data such as age, sex, and tribe were also collected. These all served as the algorithm's input variables. The data were preprocessed using the wrapper method, and all missing values were replaced using the software, which yielded better results for preprocessing.

A. Methodology

This paper aims to create a predictive diagnostic model of malaria using the decision tree classification. Confusion matrix accuracy, sensitivity, specificities, FPr, and FNr were used to assess the model’s performance. Figure 2 shows the general proposed methodology, and the five phases of the framework used in this study are as follows.

1) Preprocess the data features: The data preprocessing steps occur at this phase.

2) Cross-validate: The cross-validation is performed to evaluate the proposed model.

3) Create the decision tree.

4) Save the best classifiers network.

5) Test proposed model.

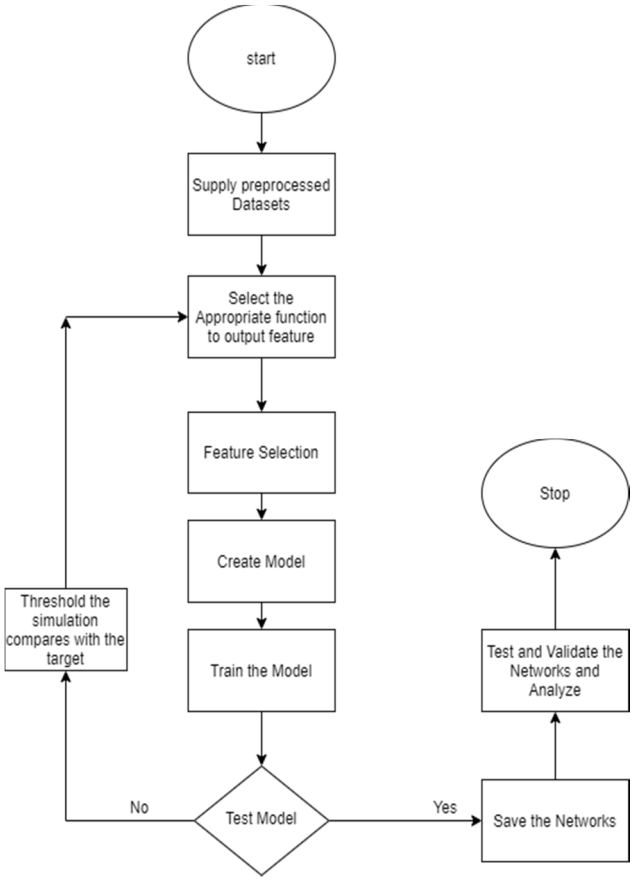

Fig.2. A framework of decision tree classifiers for malaria diagnosis

B. Features Descriptions

Table 1 represents the feature description of the model. The model used the wrapper method of filtering to select appropriate features. The dataset has one target variable (dependent) called malaria status, classified as the diagnosis response: "uncomplicated," "serious," and "none." All the other columns are the predictors (independent variables): the patient's symptoms, sex, age, and fever days. The symptoms include vomiting, high temperature, headache, and other symptoms for which the Yes/No value is classified. The data type for all variables is the type of factor.

TABLE I. FEATURES DESCRIPTIONS

Variable	Definition	Role
Malaria Status	Uncomplicated, Severe, None	Target
Age	"0–17 years," "18-25 years," "26–40," "41 & above," "0–1"	Independent
Sex	Male, Female	Independent
Fever days	"0–4 day," "5–7 days," "8–10," "11 days & above."	Independent
High temperature	Yes, No	Independent
Headache	Yes, No	Independent
Cough	Yes, No	Independent
Vomit	Yes, No	Independent
Weakness	Yes, No	Independent
Sweat	Yes, No	Independent
Loss of Appetite	Yes, No	Independent
Skin rash	Yes, No	Independent
Abdominal Pain	Yes, No	Independent
Constipation	Yes, No	Independent
Convulsion	Yes, No	Independent
Diarrhea	Yes, No	Independent
Nausea	Yes, No	Independent
Frequent-Urination	Yes, No	Independent
Muscle Pains	Yes, No	Independent
Headache	Yes, No	Independent
Cough	Yes, No	Independent
Vomit	Yes, No	Independent
Weakness	Yes, No	Independent
Sweat	Yes, No	Independent
Loss of Appetite	Yes, No	Independent
Skin rash	Yes, No	Independent
Abdominal Pain	Yes, No	Independent
Constipation	Yes, No	Independent
Convulsion	Yes, No	Independent
Diarrhea	Yes, No	Independent
Nausea	Yes, No	Independent
Frequent-Urination	Yes, No	Independent
Muscle Pains	Yes, No	Independent

The cases (diseases) are the column that has been transformed into the target variable, where severe plasmodiasis R/O CSM, severe malaria2dT yp, severe malaria knows HTN/DM, severe malaria + severe UTI, severe malaria + BPH, etc., have all been classified as "Severe." Malaria in the column has been classified as "uncomplicated," and other issues are classified as "none," meaning the patient. Other diseases, such as typhoid, share malaria symptoms, but they are different. The target has been transformed into codes: "severe" = "0," "uncomplicated" = "1" and "none" = "2." The age has been transformed into "1–17 years" = "0" ranges, and the column for sex was transformed into codes: "1" for male and "0" for female. The fever days simply refer to the days the patient felt feverish before reporting to the health center. These days have also been transformed into "0" for "0–4 days," "1" for "5–7 days," "3" for "8–10 days" and "11 and above" as "3." The associated symptoms were divided into various columns with a symptom column. The researcher then recorded yes or no under the symptoms as they appear in the datasets: the first row will have a yes for high temperature, yes for headache, yes for cough, yes for vomiting and anemia, whereas any other symptoms not shown in this row will have no value.

C. *Exploratory Data Analysis*

The data need to be visually explored to understand the distribution of individual variables/features, find missing values, and determine their relationship with other variables. Let us start with the univariate explanatory data analysis (EDA) that involves individually exploring variables. Using the bar plots, Figure 3 shows the frequency of malaria status with "None," "Severe," and "Uncomplicated," while Figure 4 shows the malaria status based on age group and sex. This will be used to visualize the categorical variables that are part of the data exploration.

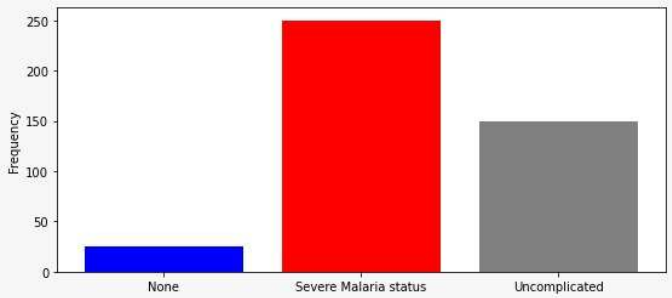

Fig. 3. Frequency of malaria status with None, Severe, and Uncomplicated status

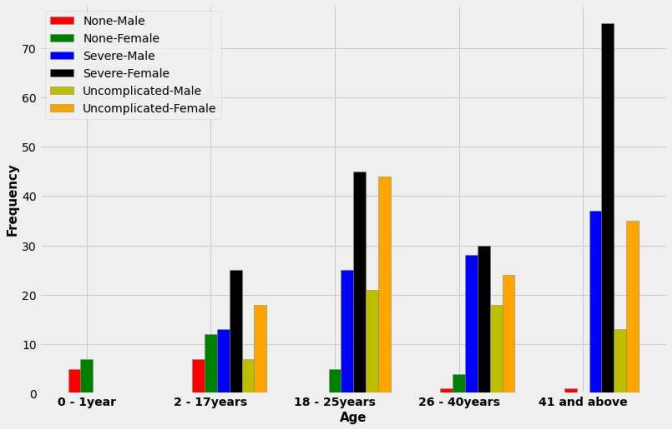

Fig. 4. Frequency of malaria status based on Sex and Age group

Figure 4 shows that most of the patients diagnosed are old. It also implies that young people mostly refuse to go for medical check-ups because they have a lower ratio. It has also been found that females go for a medical diagnosis more often than males. Sokoto has a high rate of analphabetism, and males are thought to be working and maybe so busy that they might not have time to check.

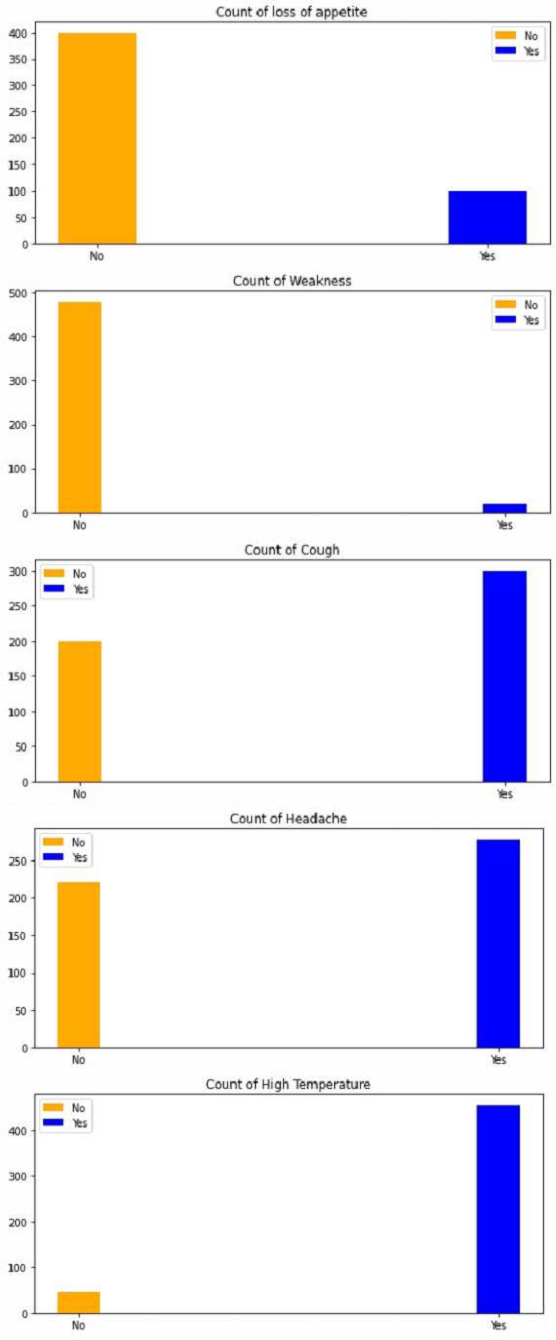

Fig. 5. Frequency of diagnosis between age groups and their gender

Figure 5 shows the visual analysis of some of the datasets; the most common symptoms among patients diagnosed at the hospital are headache, cough, and high temperature. Finally, the researcher found that most patients do not report fever at an early stage. The majority report fever at about eight to ten days, and thus this disease might progress to the severe stage and lead to the death of patients.

IV. Model Building

The decision tree in Figure 6 uses the classification and regression trees (CART) algorithm and certain forms of measures to divide into nodes and branches. Out of the 38 independent variables used to build the predictive model, age appears to have the greatest predictive power, which was measured using the Gini and entropy. After age, headache also has an impact on the prediction and difficulty in breathing. The decision tree seen can be used to group patients who will come to the hospital in the future; patients and health workers will be able to tell with a certain likelihood if a patient will belong to a malaria status class.

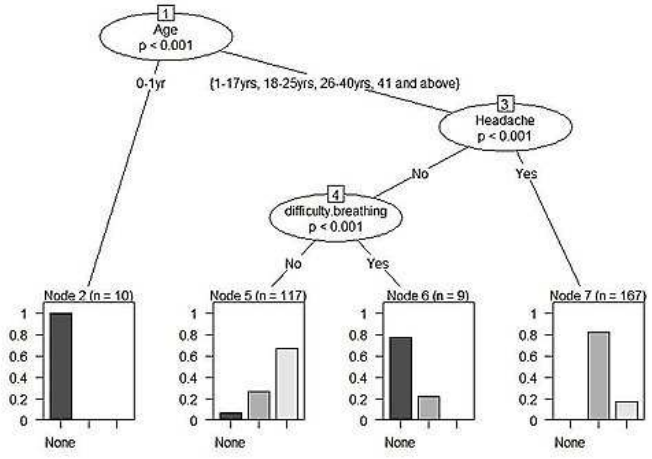

Fig. 6. CART Algorithm Decision Tree

Decision trees produce a set of rules for generating predictions for a new set of data. In this work, several models of decision trees were developed to classify patients into the three stages of malaria using notable algorithms. It has been shown that data mining tools in the health sector can largely be used to predict a patient's malaria status to be diagnosed in the future. The rpart and ctree packages of R programming produced the highest classification rate of 77%.

More than half of the independent attributes do not contribute to the model produced for various algorithms. Of the 38 independent attributes considered in the construction of the predictive model, the three decision trees generated are implemented using different algorithms, which means that the results may differ slightly. It has been found, however, that some common attributes have contributed to all the models generated, including age, difficulty breathing, headache, and anemia.

The predictive model generated using rpart and ctree implies that, in the future, if a patient over one year of age has a headache with difficulty breathing, then that patient is said to have an 80% chance of being infected with other diseases aside from malaria while having 20% chance of being classified at the early stage of malaria. Also, it is said that patients below age one are 100% likely to have a disease other than malaria. Such a patient is expected to have a higher chance of being classified as having the severe stage of the deadly disease for patients over one year of age with only a headache.

A. Model Accuracy

The performance of the models was analyzed as shown in Figure 7 using the metrics of accuracy, false positives, and false negatives in equations

$$Accuracy = \frac{correct\ classified\ pattern}{total\ patterns}$$

$$Classification\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100$$

TABLE II. Model Confusion Matrix

	None	Severe	Uncomplicated
None	5	1	1
Severe	4	86	22
Uncomplicated	6	21	51

V. Conclusion

The researcher splits the dataset into 60:30:10 for the respective training, testing and validation dataset, built a predictive model in R using packages called the "party" and "rpart," and developed a decision tree with an accuracy of 77% while having a margin of error of 23% for both packages. The researcher discovered both algorithms had the same accuracy and produced a predictive model.

Based on the algorithms, the attributes with the highest effect on the model were given, as they have different splitting measures. Under any measure of splitting the algorithm, the root node in all the trees generated is said to have the greatest influence. For the ctree algorithm, age contributed more to the predictive malaria model, followed by headache and then the difficulty in breathing.

The conclusion, according to the analysis, is that most women in old age are more likely to have the severe stage of malaria, which implies the death of many elderly women; people still believe, however, that malaria primarily kills children under five years of age, as most researchers have pointed out in recent years.

References

[1] WHO, "Malaria microscopy quality assurance manual-second edition," World Health Organization, 2016.

[2] WHO, "Malaria microscopy quality assurance manual-second edition," World Health Organization, 2017.

[3] WHO, "World malaria report," World Health Organization, 2016.

[4] K., Kigozi, R., Charland, K., Dorsey, G., Kamya, M., & Buckeridge, D., "Predicting malaria in a highly endemic country using environmental and clinical data," *Online Journal of Public Health Informatics*, 6(1), 2013.

[5] GÜL, S., UÇAR, M. K., ÇETİNEL, G., BERGİL, E., & BOZKURT, M. R., "Automated pre-seizure detection for epileptic patients using machine learning methods," *International Journal of Image, Graphics & Signal Processing*, vol. 9 no. 7, 2017.

[6] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I., "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015.

[7] Urquiza, J. M., Rojas, I., Pomares, H., Herrera, J., Florido, J. P., Valenzuela, O., & Cepero, M, "Computers in biology and medicine," Using machine learning techniques and genomic/proteomic information from known databases for defining relevant features for PPI classification, vol. 42 no. 6, pp 639–650, 2012.

[8] Worner, S. P., & Gevrey, M., "Modeling global insect pest species assemblages to determine risk of invasion," *Journal of Applied Ecology*, 2006.

[9] Patel, J., Shah, S., Thakkar, P., & Kotecha, K., "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning," *Expert Systems with Applications*, 2015.

[10] Zinszer, K., Kigozi, R., Charland, K., Dorsey, G., Kamya, M., & Buckeridge, D., " Predicting malaria in a highly endemic country using environmental and clinical data Sources," *Online Journal of Public Health Informatics*, 6(1), 2013.

[11] S.S, Shruti A. & Shirgan, "Automatic Diagnosis of Malaria Parasites Using Neural Network and Support Vector Machine," *International Journal of Advanced Foundation in Computer* (*IJAFRC*), vol. 2, pp. 62–65, 2015.

[12] 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*), vol. 4, no. 12, 2015.

[13] Abisoye, Opeyemi A & Jimoh Gbenga R, "Symptomatic and climatic based malaria threat detection using multilevel thresholding feedforward neural network," *I.J. Information Technology and Computer Science*, pp. 8. 40–46, 2017.

[14] Opeyemi A. Abisoye, Rasheed G. Jimoh, "Comparative study on the prediction of symptomatic and climatic based malaria parasite counts using machine learning models," *International Journal of Modern Education and Computer Science* (*IJMECS*), pp. 18–25, 2018.