Comparative Study on the Prediction of Symptomatic and Climatic based Malaria Parasite Counts Using Machine Learning Models

Opeyemi A. Abisoye
Department of Computer Science, School of Information and Communication Technology, Federal University of Technology, P.M.B. 65, Minna, Niger State, Nigeria.
Email: oabisoye@futminna.edu.ng, opeglo@yahoo.com.au

Rasheed G. Jimoh
Department of Computer Science, Faculty of Communication and Information Science, University of Ilorin, P.M.B.1515, Nigeria.
Email: jimoh_rasheed@unilorin.edu.ng, jimoh_rasheed@yahoo.com

Received: 01 December 2017; Accepted: 30 January 2018; Published: 08 April 2018

Abstract—Dynamics of Malaria parasite diagnosis is complex and been widely studied. Research is on-going on the effects of climatic variations on symptomatic malaria infection. Malaria diagnosis can be asymptotically or symptomatically low, mild and high. An analytical program is needed to detect individual malaria parasite counts from complex network of several infection counts. This study adopted the experimental malaria parasite counts collected from selected hospitals in Minna Metropolis, Niger State, Nigeria and Climatic data collected at the time the experiment was conducted from NECOP, Bosso, FUT Minna, Niger State, Nigeria. One thousand and two hundred (1,200) experimental data were collected and two classifiers Support Vector Machine (SVM), Artificial Neural Network (ANN) do the prediction. Experimental results indicated that SVM produced Accuracy 85.60%, Sensitivity 84.06%, Specificity 86.49%, False Positive Rate (FP) 0.1351% and False Negative Rate (FN) 0.1594% than Neural Network model of Accuracy 48.33%, Sensitivity 60.61%, Specificity 45.48%, low False Positive Rate (FP) 0.5442% and False Negative Rate (FN) 0.3939% as depicted in their respective confusion matrix.

Index Terms—Malaria, Prediction, Artificial Neural Network (ANN), Support Vector Machine (SVM), Symptomatic, Climatic

I. INTRODUCTION

Malaria is caused by a parasite known as Plasmodium spp being transmitted by an Anopheles mosquito [1]. The parasites invade the blood and causes adverse effect on the blood cells. Within 48 to 72 hours the parasites multiply inside the red blood cells and break open, infecting more red blood cells. The first symptoms usually occur between 10-14 days to 4 weeks after infection [2]. Malaria parasites can also be transmitted from a mother to her unborn baby (congenitally), by blood transfusions and by sharing needles used to inject drugs [3]. Malaria infection has a vast outbreak especially in tropical regions: an upsurge in the rate of avoidable deaths as well as an exponential increase in the population [4]. In some part of the world, malaria parasites have developed resistance to insecticides and antibiotics [5].

Likewise, malaria researchers are pursuing a vaccine and methods that would curb the disease for good [6].

Diagnosing asymptomatic malaria transmission is not straightforward due to the obvious lack of clinical manifestations and often sub-patient levels of parasites are undetectable by microscopy [7]. Prediction of the symptomatic nature of malaria parasite counts combined with effects of climatic conditions is also needed to enhance the diagnosis. The presence of both symptomatic and asymptomatic diagnostic measure is very vital in detecting the transmission dynamics of malaria infection. To avoid the occurrence of new malaria outbreaks in both endemic and non-endemic areas, an improve methods are needed to decrease the parasite sources of infection by active prediction and treatment of symptomatic and asymptomatic parasite carriers.

There is a huge amount of data which is hard to understand and to interpret by humans difficulty arises; a typical example is malarial incidences [8]. So the need for a machine learning method arises. Such a machine processes the data and automatically finds structures in the data, i.e. learns. The knowledge about the extracted structure can be used to solve the problem at hand. Problems being solved by machine learning methods range from classifying observations, predicting values, structuring data (e.g. clustering), compressing data, visualizing data, filtering data, selecting relevant
components from data, extracting dependencies between data components, modeling the data generating systems, constructing noise models for the observed data, integrating data from different sensors, using classification and drawing inferences [9]. Thus, machine learning focuses on prediction based on known properties learned from the trained data sets [10].

II. RELATED WORKS

The prediction approaches ranges from statistical modeling, mathematical modeling and machine learning methods [11]. Mathematical, statistical and computational engineering models are playing a most vital role in predictions and for helping make decisions.

Recently, machine learning (ML) is used in medical science to check health condition [12-14] and diagnose several diseases such as cancer [15-16].

In pharmacology ML find the right formula and reliable drugs to incapacitate a disease virus [17, 18]. ML is also used to choose the effective therapeutic treatment [19]. Also ML can also be used in agriculture to increase agricultural production as with predicting pest plants [20]. In the business world ML is used to predict the stock market and stock price index movement [21].

Malaria prediction is now being conducted in many countries and typically uses data on environmental risk factors, such as climatic conditions, to forecast malaria incidence for a specific geographic area over a certain period of time [22].

An Automatic Diagnosis of Malaria Parasites using Neural Network and Support Vector machine was proposed in 2015. Since mistakes are inevitable in manual counting diagnosis and time consuming we need to develop an image processing algorithm to automate diagnosis of malaria on thin blood smears. Morphological and novel threshold selection technique can be used to identify the parasites on microscopic slides. Behavioural image features such as colour, texture and the geometry of the cells and parasite was generated. Image processing was used to identify malaria parasite with the use of Phase of Image, Mean of Greenplane, Skewness, Kurtosis, standard deviations and energy. ANN classifier gives an accuracy of 80% for affected and 77% for not affected and SVM gives an accuracy of 90% for affected and 100% for not affected. But the researchers were unaware that the performance of a classifier depends on the domain under discussion. The research focuses on asymptomatic image processing. It does not give considerations to effects of symptomatic and climatic conditions [23].

An Automatic Detection of malaria parasites for estimating parasitemia was proposed in 2015. The motivation of the research was that most of the conventional microscopy used in diagnosis of diseases is occasionally proving in efficient and results are difficult to reproduce. Three (3) classifiers SVM, Naïve Bayes and Neural network classifier and two feature extraction techniques Discrete Wavelength Transform (DWT) and Gray-Level Co-Occurrence Matrix (GLCM) were used. The system obtained 100% accuracy of disease detection with the use of SVM classifier and 92.85% accuracy with Naïve Bayes. An accuracy of 92% and 85.41% were obtained when DWT and GLCM Feature Extraction method were used respectively with Neural Network. The methods made use of the morphological, colour and texture features of Plasmodium parasites and erythrocytes not given considerations to symptomatic nature and climatic effects [24].

Malaria Outbreak Prediction Model Using Machine Learning was proposed in 2015. Early prediction of a Malaria outbreak is the key for control of malaria morbidity. This will help various health organizations to better target medical resources to areas of greatest need. Two popular data mining classification algorithms Support Vector Machine (SVM) and Artificial Neural Network (ANN) are used for Malaria Prediction. Parameters used are average monthly rainfall, temperature, humidity, total number of positive cases, total number of Plasmodium Falciparum cases and outbreak occur in binary values Yes or No. The SVM model can predict the outbreak 15-20 days in advance. The accuracy of the prediction needs to be improved on by using more training data Also, in the model the individual positive cases needed to be considered not total number of positive cases as one of the training and testing features [25].

Applying different predicting methods to the same data, exploring the predictive ability of environmental and non-environmental variables, including transmission reducing interventions and using common forecast accuracy measures will allow malaria researchers to compare and improve models and methods, which should improve the quality of malaria prediction [26].

III. MATERIALS AND METHODS

A total of one thousand and two (1,200), sampled hospitals patients laboratory experimental data were collected together with their symptomatic characteristics. Also climatic data of the respective sample data timing from NECOP weather station, FUT Minna were also collected. These all served as input variable to the network. The data was pre-processed with wrapper method and several normalization method of min-max, standardization, divide by maximum were tested. But divide by maximum gave the optimum result for the pre-processing.

A. Methodology

The objective of this paper is to analyse and compare the performance of the two classifiers Support Vector Machine and Artificial Neural Network. The Performance of the classifiers are evaluated with accuracy, sensitivity, specificity, false positive rate (FP) and false negative rate (FN). Here is the proposed general methodology as depicted in the framework in Figure 1. The framework consists of these eight (8) phases: (i) Pre-processed the data features (ii) Perform hold out cross
validation by dividing the data features into training, testing and validation (iii) Create the SVM and ANN classifiers network (iv) train SVM, ANN classifiers network (v) save the best classifiers network (vi) Test and validate the networks with testing and validation features (vii) Compare the results of the SVM and ANN classifiers network (viii) Get the best classifier

**Fig. 1.** Framework of the Comparison of SVM and ANN classifiers for Malaria Parasite Counts Prediction

**B. Features Preprocessing**

In this research features with missing data are assigned zero. A typical example is the rainfall data features. In order to standardize the range of independent variables or data features, feature scaling in equation (1) and unitary method in equation (2) were used and the binary encoding threat classes are represented in Table 1.

**a. Feature Scaling**

\[ x' = \frac{x}{\|x\|} \]  \hspace{1cm} (1)

\[ x' = \frac{x}{x_{\text{max}}} \]  \hspace{1cm} (2)

where \( x \) is an original value, and \( x' \) is the normalized value.
C. Features Description

Table 2 represents the feature description of the model. The model used the wrapper method of filtering to select appropriate features. Thus this research features is thus restricted to five (5) predominant malarial symptoms: Headache (Hd), Fever (Fv), Dizziness (Dz), Body Pain (Bp), Vomiting (Vm) and two (2) significant climatic factors that contribute to having malaria: Temperature (Temp), Relative humidity (Rh) and Rainfall (Rf).

Table 2. Features Description

<table>
<thead>
<tr>
<th>Input Description (Malaria Demographic)</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Adult (1) Children (2)</td>
</tr>
<tr>
<td>Gender</td>
<td>Male (1) Female (2)</td>
</tr>
<tr>
<td>Headache (Hd)</td>
<td>+ve (1) -ve (0)</td>
</tr>
<tr>
<td>Fever (Fv)</td>
<td>+ve (1) -ve (0)</td>
</tr>
<tr>
<td>Dizziness (Dz)</td>
<td>+ve (1) -ve (0)</td>
</tr>
<tr>
<td>Body Pain (Bp)</td>
<td>+ve (1) -ve (0)</td>
</tr>
<tr>
<td>Vomitting (Vm)</td>
<td>+ve (1) -ve (0)</td>
</tr>
<tr>
<td>Temperature (Temp)</td>
<td>{0 ≤ Temp ≤ 32.83}</td>
</tr>
<tr>
<td>Relative Humidity (Rh)</td>
<td>{0 ≤ Rh ≤ 83.74}</td>
</tr>
<tr>
<td>Rainfall (Rf)</td>
<td>{0 ≤ Rf ≤ 0.034}</td>
</tr>
</tbody>
</table>

D. Feature Classification Techniques

Table 1 represents the multiclass encoding threat classification of malaria parasite counts. Support Vector Machine (SVM), and Artificial Neural Network (ANN) classifiers were used to classify the malaria parasite counts. SVM is a binary classifier while ANN is a multiclass classifier. Since malaria parasites counts exist in multiclass nature we introduce one-against-all algorithm to SVM to serve as multiclass classifier. Pureline, Logsig and Tansig activation functions were employed with ANN to map the input signals from input nodes to the hidden layer and produce output at the output layer of the network. Also, linear, radial basis and polynomial kernel function were employed to transfer input features to the network and get appropriate results.

IV. Experimental Results

The performance of the models were analysed using the performance metrics of accuracy, sensitivity, specificity, false positives and false negatives in equations (3-6). The result is depicted in Table 3 and Figure 2 showing Artificial Neural Network Class 0 (ANN_0) with feed forward and back-propagation algorithm produced optimal 48.33% accuracy, 60.61% Sensitivity and 45.58% Specificity, FPR 0.5442% and FNR 0.3939%. Also Support Vector Machine Class 2 (SVM_2) generates optimal 85.60% accuracy, 84.06% Sensitivity and 86.49% Specificity, FPR 0.1351%, FNR 0.1594%.

\[
\text{Accuracy} = \frac{\text{Correct Classified Patterns}}{\text{Total Patterns}} = \frac{TP+TN}{TP+TN+FP+FN} \times 100
\]

\[
\text{Sensitivity (Recall)} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negative}} \times 100 = \frac{TP}{TP+FN} \times 100
\]

\[
\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \times 100 = \frac{TN}{TN+FP} \times 100
\]

\[
\text{False Positive Rate (FPR)} = \frac{FP}{TN+FP} \times 100 = 1 - \text{Specificity}
\]

\[
\text{False Negative Rate (FNR)} = \frac{FN}{TP+FN} \times 100 = 1 - \text{Sensitivity}
\]
Comparative Study on the Prediction of Symptomatic and Climatic based Malaria Parasite Counts Using Machine Learning Models

Table 3. ANN and SVM Classifier Performance

<table>
<thead>
<tr>
<th>Performance Metrics</th>
<th>ANN_0</th>
<th>SVM_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>48.33</td>
<td>85.60</td>
</tr>
<tr>
<td>Support Vectors (double)</td>
<td>308 x 8</td>
<td>308 x 1</td>
</tr>
<tr>
<td>Language-Range Multiplier</td>
<td>58</td>
<td>96</td>
</tr>
<tr>
<td>True Positive (TP)</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>True Negative (TN)</td>
<td>84.06</td>
<td>86.49</td>
</tr>
<tr>
<td>False Positive (FP)</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>False Negative (FN)</td>
<td>8.54</td>
<td>5.49</td>
</tr>
<tr>
<td>Sensitivity (TPR)</td>
<td>60.61</td>
<td>84.06</td>
</tr>
<tr>
<td>Specificity (TNR)</td>
<td>45.58</td>
<td>86.49</td>
</tr>
<tr>
<td>False Positive Rate (FPR)</td>
<td>0.5442</td>
<td>0.1351</td>
</tr>
<tr>
<td>False Negative Rate (FNR)</td>
<td>0.4558</td>
<td>0.1594</td>
</tr>
<tr>
<td>Mean Square Error (MSE)</td>
<td>0.3939</td>
<td>0.5578</td>
</tr>
<tr>
<td>Total Positive</td>
<td>33</td>
<td>69</td>
</tr>
<tr>
<td>Total Negative</td>
<td>47</td>
<td>111</td>
</tr>
<tr>
<td>Total</td>
<td>180</td>
<td>180</td>
</tr>
</tbody>
</table>

Fig.2. ANN_0 and SVM_2 Malaria Model Classifier Performance

From Confusion Matrix in Fig.3 ANN_0 Result twenty (20) cases of Class_0 infected are correctly classified as positive. This corresponds to 11.11% of all one hundred and eighty (180) malaria cases. Similarly, sixty seven (67) cases of Class_1 and Class_2 non infected are correctly classified as negative. This corresponds to 37.22% of all malaria cases. Also, thirteen (13) cases of Class_1 and Class_2 non infected cases which correspond to 7.22% are incorrectly classified as negative. Similarly, eighty (80) cases of Class_0 infected cases are incorrectly classified as positive. This corresponds to 44.44% of all malaria cases.

Out of thirty three (33) infected cases, twenty (20) were correctly classified. This corresponds to 60.61% correctly classified while thirteen (13) cases which correspond to 39.39% were wrongly classified. Similarly, out of one hundred and forty seven (147) non infected cases only sixty seven(67) cases which corresponds to 45.58% were correctly classified as non-infected cases while eighty (80) which corresponds to 54.42% were incorrectly classified.
From Confusion Matrix in Fig.4 SVM_2 Result fifty eight (58) cases of Class_2 infected are correctly classified as positive. This corresponds to 32.22% of all one hundred and eighty (180) malaria cases. Similarly, ninety six (96) cases of Class_0 and Class_1 non infected are correctly classified as negative. This corresponds to 53.33% of all malaria cases. Also, eleven (11) cases of Class_0 and Class_1 non infected cases which correspond to 6.11% are incorrectly classified as negative. Similarly, fifteen (15) cases of Class_2 infected cases are incorrectly classified as positive. This corresponds to 8.33% of all malaria cases.

Out of sixty nine (69) infected cases, fifty eight (58) were correctly classified. This corresponds to 84.06% correctly classified while eleven (11) cases which correspond to 15.94% were wrongly classified. Similarly, out of one hundred and eleven(111) non infected cases only ninety six(96) cases which corresponds to 86.49% were correctly classified as non-infected cases while fifteen(15) which corresponds to 86.49% were incorrectly classified.

V. COMPARATIVE ANALYSIS

From Table 4, the performance of the two classifiers ANN and SVM, the following comparative differences were made:

<table>
<thead>
<tr>
<th>Methodology</th>
<th>Strength</th>
<th>Weaknesses</th>
</tr>
</thead>
</table>
| SVM         | ✓ Handles Bivariate prediction, pattern recognition, feature selection and classification  
 ✓ Handles small and large dataset well  
 ✓ Uses predefined activation function  
 ✓ Solves the problems of over-fitting by optimizing the model parameters to feature selection  
 | ✓ Handles only binary prediction, pattern recognition, classification, and regression analysis  
 ✓ It needs a 'good' kernel function.  
 ✓ Choosing appropriately hyper parameters that will allow for sufficient generalization performance |
| ANN         | ✓ It does create network to have hidden neurons  
 ✓ Handle Multivariate prediction, pattern recognition, classification, regression analysis.  
 ✓ Uses predefined activation function  
 ✓ Requiring less formal statistical training  
 ✓ | ✓ No general framework to design most suited network for particular problems  
 ✓ Threshold frequency, number of hidden layers and hidden neurons are searched in the network by trial and error  
 ✓ Greater computational burden because large parameters are needed to fit a good network structure  
 ✓ Prone to local minima  
 ✓ Over fitting often occurs because of large data to fix. |
VI. CONCLUSION

In this paper, the prediction of symptomatic and climatic based malaria infection was conducted with Artificial neural Network (ANN) and Support Vector Machine (SVM). The performance evaluation of the developed ANN and SVM Malaria model was evaluated based on the threshold metrics; accuracy, sensitivity, specificity, false positive and false negative metrics sighted in Section 3. The models were comparatively evaluated as shown in Table 3 and Figure 3. ANN performance was relatively low with 48.33% irrespective of applications of different activation functions of purelin, logsig and tansig. Linear, radial basis function and polynomial kernel functions were also employed in Support Vector Machine (SVM). But performance of SVM with radial Basis function provided good results of 85.60%. Therefore, Support vector machine can be employed by medical practitioners to predict the level of severity of an infected patient. Further research can focus on improving the performance of the model possibly with hybridized models.

REFERENCES


Comparative Study on the Prediction of Symptomatic and Climatic based Malaria Parasite Counts Using Machine Learning Models

Advanced Research in Computer Engineering & Technology (IJARCET) 4(12).


Authors’ Profiles

Abisoye Opeyemi A. was born in Ogbomoso, Oyo State, Nigeria. She attended University of Ilorin, Ilorin, Nigeria where she obtained her BSc, Msc, degree in Computer Science. She is currently a PhD. Student of the same institution. She is major in Computational Intelligence, Machine Learning, Data Mining, and Soft Computing.

She serves as a Lecturer I, in the Department of Computer Science, SICT, Federal University of Technology, Minna, Niger State, Nigeria from May 23rd 2007 Till Date.

Professional Membership: (MPCN) A member of Computer Professionals[Registration Council of Nigeria]-CPN (30th June, 2010)

Jimoh Rasheed Gbenga is currently an Acting Dean of Faculty of Communication and Information Science (FCIS), University of Ilorin, Nigeria. He attended Universiti Utara Malaysia, Malaysia where he got Ph.D. in Information Technology. His research interests are: Information Security, Soft Computing and Machine Learning.

Professional Membership:

A member of Computer Professionals[Registration Council of Nigeria]-CPN
A member of Nigeria Computer Society of Nigeria (NCS).
A member of IEEE Nigeria Chapter.