A Logistic Regression Model to Predict Malaria Severity in Children

Mary Opokua Ansong, Asare Yaw Obeng, and Samuel King Opoku*

ABSTRACT

One of the main causes of death around the globe is malaria. Researchers have sought to develop predicting models for malaria outbreaks based on metrological data, climate data and the breeding cycle of plasmodium, the causative agent of malaria. This study predicts the severity of malaria based on environmental and biological factors. A logistic regression model was developed in this study to predict the severity of malaria based on such factors as sickle cell disease, stagnant water, garbage dumps, wet lawns, and the use of treated mosquito nets with an 83.3% accuracy rate. The study was carried out in the Bosomtwe District of Ghana with 417 respondents. It was deduced that although children in the district are highly prone to malaria infection, the severity is very low. The study recommends that not just having a good sample size alone is important during machine learning model development but also having a good sample representation of the various class labels is equally important.

Keywords: Bosomtwe District, Children, Logistic Regression, Malaria Severity.

1. Introduction

The causative agent of Malaria is a protozoan parasite called Plasmodium. The deadly type of the Plasmodium parasite is the Plasmodium falciparum causing about 90% fatalities of malaria in humans in Africa. Malaria has become a public health issue in Africa with the fatality rate increasing exponentially in children under five years old [1],[2]. In 2019, half of the global cases of malaria amounting to about 229 million came from six countries in Africa with Nigeria contributing to 25% of the global count [3]. The initial clinical symptom, usually fever, is followed by vomiting, tiredness, abdominal pain and diarrhoea. Failure to treat falciparum malaria within 24 hours after observing the initial clinical symptom can be fatal. Symptoms can be complicated to the extent of organ or system failure. One of the common systems usually attacked by malaria is the central nervous system resulting in cerebral malaria [3],[4].

Malaria is transmitted to humans by infected female Anopheles mosquitoes as the male Anopheles mosquitoes do not suck blood [3]. Malaria can also be transmitted from a pregnant mother to her baby, through blood transfusion and sharing of needles used to inject drugs [6]. Fig. 1 illustrates the transmission cycle of malaria.

From Fig. 1, the upper section, labelled 5, shows the development of the plasmodium parasite from gametocytes to sporozoites which can affect the human host in mosquitoes. The lower section, labelled 1 to 4, illustrates how the human host is affected. An infected mosquito bites the host and transmits the sporozoites of the Plasmodium, the malaria parasite, into the host. The parasite travels to the liver of the host where it lies dormant for about ten to twenty-eight days. In the liver, the sporozoites develop thousands of merozoites. The merozoites leave the liver and infect red blood cells. This is when malaria signs and symptoms develop. The merozoites develop into further stages called gametocytes. A mosquito bites an infected person and it is infected with gametocytes. The transmission cycle continues from point 5 in Fig. 1.

There have been various studies on malaria prediction [7]–[11]. However, the techniques of data collection and analysis of the illness prevalence and breadth varied greatly. Some of the methods employed in malaria predictions are Naive Bayes [7], [12], Support Vector Machine (SVM) [7], [12], Linear Regression [7], Logistic Regression [7], K-Nearest Neighbor (KNN) [7], [12], Random Forests [12] deep learning [13], multi-layer perceptron and decision tree, [10]. Unfortunately, predicting malaria outbreaks using linear regression has not been successful [7]. None of
the existing works has predicted the severity of malaria in children.

Researchers train their machine-learning models using data with various attributes. Some of the researchers predicted malaria outbreaks using meteorological and malaria incident data [7], meteorological data and the malaria-carrying vector (mosquito) breeding environment [13], climatic conditions [10], biological characteristics and social determinants associated with demographic and health survey [12]. None of the research works mentioned above predicted the severity of malaria based on the combination of environmental and biological factors.

This study used a logistic regression model to predict the level of malaria severity in children using environmental conditions and biological factors. Different methodologies have their own set of advantages and disadvantages. Logistic regression facilitates the conversion of complex data into simple and relevant insights. By utilizing algorithms to answer queries and have conversations, logistic regression aims to imitate human-like responses [14]. Logistic regression can be used in the healthcare sector to summarize large narrative texts, such as academic journal articles or clinical notes, by highlighting essential concepts or phrases in the reference document. To improve clinical decision-making, logistic regression can map data pieces in Electronic Health Records that are available as unstructured text into structured useful data [15]. The remaining part of this paper is divided into three sections that focus respectively on the materials and methods, results and conclusion.

2. MATERIALS AND METHODS

The model to predict the severity of malaria in children was developed using a logistic regression algorithm. This algorithm requires that certain conditions are satisfied to develop a good model. The logistic regression algorithm ensures that the dependent variable is binary with the value 1 representing the desired outcome. Certain independent variables which are not meaningful and depend on other independent variables (collinearity) should be avoided. A serious limitation of the logistic regression algorithm is associated with overfitting where the model learns both the data and the associated noise so well that it is not able to perform well with unseen data. Logistic regression requires an appreciable size of data for training. This section describes the various methods that were carried out to develop a good model. It also explains how the conditions required by the logistic regression algorithm were handled.

2.1. Data Collection and Data Processing

The study population was narrowed down to Amakom in the Bosomtwi District of Ghana to determine the severity of malaria in children under five years due to the high occurrence of malaria in the area [16]. The study population is two thousand. A sample size of four hundred and seventeen (417) children using the Stratified sampling method was ideal for the study [17]. The dataset was examined, cleaned and processed for analysis and machine learning model development.

2.2. Problem Formulation

Given the features or attributes of the independent variables $F = (f_1, f_2, f_3, ..., f_T)$ where $T$ is the number of features or attributes and the corresponding severity of malaria $S \in \{0, 1\}$ where 0 means no severe malaria and 1 means severe malaria. We determine a logistic regression function $LR(f)$ such that the predicted output $LR(f_i)$ are possibly close to the actual output $S_i$ for each observation $i = 1, 2, \ldots, n$ where $n$ is the number of observation. Given that $S_i \in \{0, 1\} \forall LR(f_i) \approx 0|1|$, the sigmoid function $\sigma(x) = \frac{1}{1 + \exp(-x)}$ was used with the logistic regression function.
Since a logistic function is a linear classifier, it implies:

\[ f(x) = \sum_{j=0}^{T} b_j x_j \text{ such that } \forall x_0 = 1 \]

The \( b_j \text{ such that } \forall j = 0, 1, 2, \ldots, T \) are the predicted weights which are the estimated coefficients. The logistic regression function \( LR(f) \) is a sigmoid function given as:

\[ LR(f) = \frac{1}{1 + \exp(-f(x))} \approx 0|1 \]

The predicted weights \( (b_j \forall j = 0, 1, 2, \ldots, T) \) are best obtained by maximizing the log-likelihood function (LF) for all the observations as:

\[ LF = \sum_i (S_i \log (LR(f_i)) + (1 - S_i) \log (1 - LR(f_i))) \]

The calculation of the best weights using the LF was handled by the logistic regression Python libraries. There are such open-source packages in Python as NumPy for manipulating arrays and Matplotlib to visualize results. The scikit-learn and statsModels are the libraries used to create, fit (or train), evaluate (or test) and apply a logistic regression model. To avoid multicollinearity, the correlation between the independent variables was determined.

### 2.3. Handling Overfitting of Logistic Regression Algorithm

One way to control overfitting is by regularizing the predicted weights. There are three methods of penalizing large coefficients in logistic regression to control overfitting. The first method, usually called L1, uses the absolute values of the predicted weights such that the linear function becomes:

\[ f(x) = \sum_{j=0}^{T} |b_j x_j \text{ such that } \forall x_0 = 1 \]

Another method, usually called L2, uses the squared values of the predicted weights so that the linear function becomes:

\[ f(x) = \sum_{j=0}^{T} b_j^2 x_j \text{ such that } \forall x_0 = 1 \]

The final method, usually called elastic-net, combines the previous two methods and the linear function becomes:

\[ f(x) = \sum_{j=0}^{T} |b_j|^2 x_j \text{ such that } \forall x_0 = 1 \]

Scikit-learn library allows regularization setting using the penalty parameter. The scikit-learn library was therefore chosen over statsModels. The scikit-learn library was used with the logistic regression class implementing the ‘liblinear’ library which requires regularization to work using L1, L2 or elasticnet as values to the penalty parameter [18]. The default penalty is L2 which was used as demonstrated in some research works [19], [20].

### 3. Results

This section presents the analysis of data collected from 417 self-administered questionnaires. These were parents who completed the questionnaire on behalf of their children under five years. The findings were presented in tables and figures to complement the interpretation of the data collected.

#### 3.1. Descriptive Statistics of the Data Collected

This section describes the features of the data by generating summaries of the data collected. Table I describes how often respondents’ children get malaria every year.

Table I shows that 204 (48.9%) of the respondents’ children often do not get malaria every year, though the respondents’ children had contracted malaria before but it does not occur every year. 213 (51.1%) of the respondents’ children contract malaria from one to more than five times a year. This shows how prevalent malaria occurs in the area.

Table II shows the severity of malaria in respondents’ children. Respondents were asked to describe the severity of malaria that affects their children.

Table II shows that 336 (80.6%) of malaria occurrence is not severe. These children do not stay in clinics or hospitals for more than 24 hours. However, 81 (19.4%) of malaria occurrences are severe or fatal. These children usually stay in hospitals for more than 24 hours and they are placed in intensive care units.

Table III summarises the environmental conditions of the respondents. It shows the number of respondents answering Yes and those answering No for the selected environmental conditions.

From Table III, the environmental conditions considered in this study were the presence of mosquitoes in the area. Respondents answered this question based on mosquito sounds they hear and whether they see them physically. Other environmental conditions looked at the nature of the environment, namely, the occurrence of wet lawns in the respondents’ residence or close to the residence; refuse dump within 300 meters of the respondents’ area.

---

**TABLE I: HOW OFTEN CHILDREN GOT MALARIA**

<table>
<thead>
<tr>
<th>Often malaria</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>No malaria</td>
<td>204</td>
<td>48.920863</td>
</tr>
<tr>
<td>Up to 2 times</td>
<td>157</td>
<td>37.649880</td>
</tr>
<tr>
<td>3 to 5 times</td>
<td>33</td>
<td>7.913669</td>
</tr>
<tr>
<td>More than 5 times</td>
<td>23</td>
<td>5.515588</td>
</tr>
<tr>
<td>Total</td>
<td>417</td>
<td>100</td>
</tr>
</tbody>
</table>

**TABLE II: SEVERITY OF MALARIA**

<table>
<thead>
<tr>
<th>Severity</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not severe</td>
<td>336</td>
<td>80.58</td>
</tr>
<tr>
<td>Severe</td>
<td>48</td>
<td>11.51</td>
</tr>
<tr>
<td>Very severe</td>
<td>18</td>
<td>4.32</td>
</tr>
<tr>
<td>Resulted in death</td>
<td>15</td>
<td>3.60</td>
</tr>
<tr>
<td>Total</td>
<td>417</td>
<td>100.0</td>
</tr>
</tbody>
</table>
A Logistic Regression Model to Predict Malaria Severity in Children

Ansong et al.

### TABLE III: ENVIROMENTAL CONDITIONS

<table>
<thead>
<tr>
<th>Answer</th>
<th>A lot of mosquito</th>
<th>Sleep under mosquito net</th>
<th>Wet lawns in the residence</th>
<th>Refuse dump within 300 m</th>
<th>Constant Stagnant water around the residence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>261</td>
<td>250</td>
<td>218</td>
<td>208</td>
<td>267</td>
</tr>
<tr>
<td>No</td>
<td>156</td>
<td>167</td>
<td>199</td>
<td>209</td>
<td>150</td>
</tr>
<tr>
<td>Total</td>
<td>417</td>
<td>417</td>
<td>417</td>
<td>417</td>
<td>417</td>
</tr>
</tbody>
</table>

### TABLE IV: SICKLE CELL ANEMIA GENOTYPE OF CHILDREN

<table>
<thead>
<tr>
<th>Sickle cell anemia</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>203</td>
<td>48.68</td>
</tr>
<tr>
<td>No</td>
<td>214</td>
<td>51.32</td>
</tr>
<tr>
<td>Total</td>
<td>417</td>
<td>100</td>
</tr>
</tbody>
</table>

### TABLE V: CORRELATION RESULTS

<table>
<thead>
<tr>
<th></th>
<th>SC</th>
<th>OM</th>
<th>SW</th>
<th>RD</th>
<th>WL</th>
<th>M</th>
<th>SN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sickle cell (SC)</td>
<td>0.1</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Often malaria (OM)</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>Stagnant water (SW)</td>
<td>0.3</td>
<td>0.1</td>
<td>0.4</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>Refuse dump (RD)</td>
<td>0.2</td>
<td>0.1</td>
<td>0.4</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>Wet lawns (WL)</td>
<td>0.2</td>
<td>0.1</td>
<td>0.3</td>
<td>0.3</td>
<td>0.4</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Mosquitoes (M)</td>
<td>0.1</td>
<td>0.1</td>
<td>0.3</td>
<td>0.3</td>
<td>0.4</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>Sleep in net (SN)</td>
<td>0.1</td>
<td>0.1</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.2</td>
<td></td>
</tr>
</tbody>
</table>

### TABLE VI: CLASSIFICATION REPORT OF TEST DATA

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decrease severity of malaria (No)</td>
<td>0.88</td>
<td>0.93</td>
<td>0.90</td>
<td>70</td>
</tr>
<tr>
<td>Increase severity of malaria (Yes)</td>
<td>0.50</td>
<td>0.36</td>
<td>0.42</td>
<td>14</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.83</td>
<td>0.84</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A biological factor which focuses on the genotype of sickle cell Anemia in respondent children was described in Table IV. Table IV shows that 203 (48.68%) were sickle cell anaemic or carriers. These children have S-gene as either SS or AS genotypes. However, 214 (51.32%) were considered non-sickle cell anaemic. They do not have S-gene, though they may have the C-gene. They are considered AA or AC genotypes.

### 3.2. Collinearity of Independent Variables

To ensure that the independent variables do not depend on themselves and avoid multicollinearity, a correlation matrix between the independent variables was calculated. Table V shows the correlation results.

From Table V, the results of correlation among the factors do not have any effect on each other variables. None of the correlation scores is above 0.5. This suggests that the variables will serve to give a meaningful explanation of the model to be developed in this study.

### 3.3. Logistic Regression Results

Out of 417 records obtained, 333 records were used for training and 84 for testing. Table VI shows the classification report when the model was tested with the test data.

From Table VI, the precision examined each side individually, identifying 0.88 (88%) No correct on the No side and 0.50 (50%) Yes correct on the Yes side. The Recall determined that 0.36 (36%) correct Yes and 0.93 (93%) correct No were on either side of the Yes/No. The support is identical to the confusion matrix depicted in Table VII.

Table VII summarizes the performance of the model on the test data, based on the confusion matrix. 5 respondents said their children had severe malaria and the prediction was the same (true positive). 65 respondents said their children did not have severe malaria and the prediction was correct (true negative). 5 respondents said their children had severe malaria but the algorithm says no (false negative). 9 respondents said their children do not have severe malaria but the algorithm says yes (false positive). The outcome shows that we made 70 (65 + 5) correct predictions. According to the outcome, we made 14 (9 + 5) incorrect predictions. The accuracy rate is therefore 83%.

### 4. Conclusion

A logistic regression model was developed in this study to predict the severity of malaria in children under five years old in Amakom of Bosomtwe District in Ghana using parameters from environmental and biological backgrounds with an accuracy rate of 83%. The model performs poorly when predicting 1 (that is yes) to indicate the severity of malaria with 42% accuracy taking into consideration false positives and false negatives. This was due to the low representation of data on malaria severity. However, the model performs creditably well when predicting 0 (that is no) to indicate non-severity of malaria with 90% accuracy taking into account false positives and false negatives. It can therefore be deduced that although children in Amakom are highly prone to malaria infection, the severity is very low. Moreover, not just having a good sample size alone is important during machine learning model development but also having a good sample representation of the various class labels is equally important. Future work should focus on extending the work to cover the whole country to have an appreciable number of respondents for 1 (that is yes) to indicate the severity of malaria.

### Acknowledgment

We wish to thank all the respondents who contributed to the success of this study by completing the questionnaire.
**Conflict of Interest**

The authors declare that they do not have any conflict of interest.

**References**


