PREDICTIVE MODELING OF MALARIA DRUG RESISTANCE CASE STUDY GURATOPP RAYFIELD PRIMARY HEALTHCARE, PLATEAU STATE, NIGERIA

Jamilu Yunusa Falgore*, Alkali Ivy Iveren, Alkali Msoo and Jinad Ololorin Sodiq

Department of Statistics, Ahmadu Bello University, Zaria - Nigeria

*Corresponding author: *jamiluyf@gmail.com*

ABSTRACT

The primary approach for treating malaria has been Artemisinin-based Combination Therapy (ACT), yet emerging resistance threatens its effectiveness. This study applies a logistic regression framework to predict malaria drug resistance among patients at Guratopp Rayfield Primary Healthcare Center. Using a data set of 500 patient records, predictors included age, type of diagnostic test, test result, drug type and dosage administered. The study explores the relationship between patients and resistance to malaria treatment, as well as, building a binary classification model on drug resistance among patients. From the malaria tests conducted, results showed that drug dosage and positive malaria test outcomes, were positively associated with resistance, while test type and negative results, were significant predictors of non-resistance. Logistic regression and GridSearch cross-validation both produced an accuracy score of 74% in classifying resistance status, outperforming alternative models. The model coefficients estimated via a maximum likelihood, highlight the statistical relationship between patient characteristics and resistance outcomes. Findings emphasize the importance of statistical modeling in malaria research, providing a replicable framework for monitoring resistance patterns. This approach offers healthcare providers and policymakers data-driven insights for adapting treatment protocols and enhancing surveillance of drug resistance.

ISSN NUMBER: 1116-249X

Keywords: malaria treatment, binary classification, drug resistance, machine learning, logistic regression

1. INTRODUCTION

Malaria is a multifaceted disease whose epidemiology and clinical features differ considerably across regions. These variations are influenced by factors such as the prevailing Plasmodium species, their sensitivity to available antimalarial drugs, the abundance and effectiveness of mosquito vectors, environmental and climatic conditions, and the immunity and behavioral patterns of local populations (WHO, 2022). Although several strategies have been developed to control malaria, no single approach is universally suitable or affordable. Effective prevention and control measures must therefore be adapted to the specific context, accounting for local epidemiology, available resources, and political commitment. Antimalarial drug resistance has further complicated control efforts, contributing to malaria's spread into new areas, its re-emergence in previously cleared regions, and the increased severity of epidemics in some settings. The increasing complexity of these challenges has made the traditional methods of surveillance and control less effective, necessitating innovative approaches such as data-driven methodologies and machine learning models to predict, track, and analyze malaria resistance patterns with greater precision (WHO, 2022).

Malaria transmission is most common in warm and humid regions, particularly across sub-North Africa, Southern and Central United States, Hispaniola in the Caribbean, the Middle East, the Indian subcontinent, Southeast Asia, and parts of Oceania (World Health Organization, 2022). Despite its wide distribution, transmission intensity and infection risk vary greatly between regions. For instance, highland areas above 1500 meters and arid regions with less than 1000 mm

ISSN NUMBER: 1116-249X

of annual rainfall tend to have lower transmission rates, yet they remain vulnerable to epidemics when infected individuals are present and climatic conditions favor mosquito breeding (WHO, 2022). Advanced computational approaches such as machine learning can now be employed to analyze these variations and predict malaria hotspots by processing complex epidemiological and environmental data, providing invaluable insights for effective intervention strategies.

Malaria has been intensely researched as one of the greatest persistent infectious disease caused by parasites in human history. In 2021, approximately 247 million cases were reported across 84 endemic countries, a slight rise from 245 million in 2020, with the majority of new cases occurring in the African Region of the WHO (WHO, 2022). At the start of the Worldwide Technical Plan for Malaria 2016–2030, in 2015, the global case estimate was about 230 million. Frequency of cases fell from 82 for every 1000 people in 2000, there were individuals at risk and that number increased to 57 by 2019, but increased again to 59 in 2020, largely because of interruptions caused by the COVID-19 outbreak. These service interruptions are thought to have contributed an additional 13.4 million cases between 2019 and 2021 (WHO, 2022). The share of malaria instances resulted from Plasmodium vivax also declined, from 8% (20.5 million cases) in the year 2000, the percentage dropped to 2% (4.9 million) by 2021. A total of twenty nine countries were responsible for the majority of malaria cases worldwide (96%), with Nigeria accounting for 27%, the Democratic Republic of Congo for 12%, Uganda for 5%, and Mozambique for 4%. together, they represent a significant portion of the global total. In 2021, the African Region of WHO alone recorded around 234 million instances, representing about 95% of the worldwide burden. The ongoing issue of drug resistance continues to pose significant obstacles to the eradication of malaria. Though plasmodium falciparum accounts for the majority of fatalities related to malaria, Plasmodium vivax has also emerged as an important cause of severe disease and mortality (WHO,

ISSN NUMBER: 1116-249X

2013). Machine learning and AI-driven approaches are gaining prominence in analyzing these trends to serve as predictors of emerging outbreaks, especially as climate change and human movement influence malaria dynamics.

For many years, resistance to antimalarial medications has posed a significant challenge to efforts in fighting malaria. The purpose of malaria drugs is to eliminate the malaria parasites present in the bloodstream of infected individuals, thereby reducing potential sources of infection within the community. Drug resistance, which refers to the ability of a parasite strain to survive or reproduce even when the host has taken therapeutic doses of the medication safely, continues to be a significant barrier in malaria control. Traditionally, chloroquine was the main treatment option for malaria, but the emergence of chloroquine resistance throughout endemic regions in the 1980s led to modifications in treatment protocols that favored the use of Artemisinin-based combination therapy (ACT) as the preferred medication. Over the past decade, ACT has been instrumental in decreasing the mortality rate associated with malaria (Aribodor et al., 2016). However, predictive models leveraging machine learning can now assist in monitoring resistance trends and even anticipate potential drug resistance hotspots, giving researchers and policymakers the capacity to act proactively.

Up till now, Drug resistance has been observed in three out of the five species of Plasmodium, namely P. falciparum, P. vivax, and P. malariae, which are the pathogens that cause malaria in humans (Sinha et al., 2014). The extent of malaria concerning its effect on disease and death among humans positions it as a significant public health issue in tropical and subtropical regions. Nevertheless, in spite of the remarkable early successes of the National Malaria Control and Eradication Programs launched in the 1950s, In numerous countries, attempts to eradicate malaria were entirely unsuccessful owing due to challenges related to technical, operational, and socio-

ISSN NUMBER: 1116-249X

economic factors, which cause a revival of malaria across numerous regions of the globe. The spread of drug resistance in parasites and insecticide resistance in mosquito vectors has compromised the efficiency of control programs (Borne, 2004). Machine learning techniques can offer valuable insights into how environmental factors, parasite resistance, and transmission are interconnected, aiding in the development of more intelligent and sustainable control methods. Drug resistance frequently leads to a delayed or ineffective elimination of asexual parasites from the bloodstream, which ultimately contributes to the generation of gametocytes, the parasites that transmit the resistant genotype. Following the WHO's endorsement of artemisinin-based combination therapies (ACTs) as the primary treatment for P. falciparum malaria in 2001 (WHO, 2001), there was a notable reduction in malaria outbreaks after 2005. Nonetheless, drug-resistant variants of artemisinin and its derivatives have recently surfaced in various parts of Southeast Asia, threatening the advancements made in malaria control, treatment, and elimination efforts (Sinha, 2014).

1.1. Causative Agents of Malaria

Malaria in humans results from infection by one or more of four intracellular protozoan parasite species, with Plasmodium being the causative agent, which are spread to people by means of bites from infected female Anopheles mosquitoes. Numerous species of Plasmodium exist, including Plasmodium falciparum, P.vivax, P.ovale, and P.malariae (however, P. falciparum is the most common species found in sub-Saharan Africa (Chinedu O.E et al., 2023)), vary in geographical distribution, microscopic morphology, and clinical characteristics, including infection periodicity, risk of serious illness, and capacity to trigger recurrences), and their capacity to acquire resistance to antimalarial drugs. So far, drug resistance has become an issue observed in just two of the four species, P.falciparum and P.vivax.

ISSN NUMBER: 1116-249X

1.2. Antimalarial Drug Resistance

Antimalarial drug resistance refers to the capacity of a strain of a parasite to endure or reproduce even when a medication is given at standard or higher doses, provided the host tolerates it. The The definition was subsequently updated to clarify that the medication must access the parasite or the infected red blood cell for the necessary duration required for its effect (WHO, 2001). In areas where transmission is consistently high, 70 host immunity plays a significant role in clearing partially resistant parasites and allows older children and adults can harbor significant parasite loads without showing symptoms (Sarda et al., 2009; Lopera-Mesa et al., 2013). These asymptomatic carriers dilute the population-level influence of antimalarial medications (White, 2004). Regions with high transmission also face infections from multiple strains, where genetic recombination in mosquitoes increases parasite diversity and competition, reducing the fixation of resistant alleles (Jiang et al., 2011; Takala-Harrison & Laufer, 2015). Conversely, low-transmission regions have fewer multiple infections, lower immunity, and more symptomatic patients, who may receive substandard drugs, incomplete courses, or monotherapies, facilitating the emergence of resistance. Populations of P. falciparum in Southeast Asia exhibit significant structural organization, with genetic backgrounds that predispose parasites to develop resistance through multi-stage processes (Miotto et al., 2013). Following WHO's recommendation of artemisinin-based combination therapies (ACTs) serve as the primary treatment for P. falciparum malaria, outbreaks declined substantially (WHO, 2001). Nonetheless, artemisinin-resistant parasites have emerged, highlighting the need for continued analysis of antimalarial drug resistance.

This study aimed to analyze malaria drug resistance with a focus on identifying patient and treatment related factors such as age, drug dosage, diagnostic test type and test results that may

ISSN NUMBER: 1116-249X

influence resistance. Furthermore, the study applied logistic regression modelling and cross-validation techniques to statistically predict resistance outcomes among patients at the Guratopp Rayfield Primary Health Care Center.

2. LITERATURE REVIEW

A study on antimalarial drug resistance carried out by Cui et al., (2015) demonstrated the resistance level of anti-malarial medications. Malaria parasites have been found to be able to develop resistance to every new class of medications in an evolutionary arms race with various therapeutic interventions. They suggested that more anti-malaria drugs should be provided to public health centers and awareness programs should be made. Data was collected from ICEMR (International Centers of Excellence for Malaria Research) sites. To achieve this, analysis of variance and logistic regression techniques were used to analyze the dataset using SPSS software.

Ndiath et al. (2015) employed regression techniques to identify risk factors associated to areas of high malaria transmission within the Keur Soce surveillance zone. The study also compared models using ordinary least squares. It investigated the geographical correlations between the prevalence of malaria and environmental and socioeconomic factors in tiny settlements in Keur Soce, utilizing data from the International Institutes of Excellence in Malaria Research (ICEMR) sites. Data analysis was conducted with analysis of variance and logistic regression using SPSS software, based on six months of passive surveillance in Senegal. Blood film for the “gold standard” microscopy verification of the Rapid Diagnostic Tests (RDTs) test was carried out and determinant factors at the household level were examined using an individual questionnaire. A software ArcMap version 9.2. Was employed for mapping visualization, computing and exploratory analysis. According to the study, high numbers in situations of malaria cases are

ISSN NUMBER: 1116-249X

primarily concentrated in the southeastern region where hotspots are more precise, but smaller values are predominant in the center also within the north.

Boateng et al., (2015) analyzed using Quasi-Poisson Regression to determine malaria prevalence in Obuasi Municipality, Ghana. The data management department of the Obuasi Government Hospital in Ghana collected the malaria data from the Out Patients Department (OPD) between January 2007 and December 2010. According to the regression statistical model, variables, results, effect estimate, and restrictions, people over 50 and those between the ages of 20 and 34 have a disproportionately high incidence of malaria. When wide phrases were used to search the PubMed, Embase, and Scopus databases, 4,281.

Using a logistic regression model and maximum likelihood estimate, Dalatu et al. (2015) investigated factors linked to malaria. Fever, which is a temperature of at least 37.5 degrees, headaches, convulsions, colds, coughing or perspiration, etc. are some of the predictors that affect malaria in children and adults, both male and female, in Kebbi State metropolis. The model's application using the state's malaria data established a significant relationship between malaria status and such predictors. Based on the collected data, approximately 99.6% of the predictions were accurate. After analysing the data, the study came to the conclusion that fever, a temperature of at least 37.5 degrees, and other factors are predictors that truly affect malaria. The best way to prevent malaria, according to the scientists, is to use treated nets, particularly now that drug-resistant insects are emerging. Efforts should be made to remove conducive environments where anopheles can lay eggs, according to additional recommendations. Malaria cases can also be decreased by keeping the surroundings clean. In light of the aforementioned, the authors recommended that physicians and clinics use the models created by this study to determine the

ISSN NUMBER: 1116-249X

prevalence of malaria in individuals so that appropriate preventative and control measures can be implemented early enough to avoid the risk of the disease's full manifestation.

Black and Jackson (2017). The purpose of this study of the research was to determine whether the present version of the Lives Saved Tool (LiST) model should include the effect of malaria on stunting. The following information was taken from pertinent studies by Jackson and Black: year, country, population, design, number of subjects, age range, malaria endemicity, Plasmodium species, intervention, duration and timing, malaria episode definition, exposure, stunting reference standard, outcome, regression statistical model, covariates, results, effect estimate, and limitations. Using general search criteria, 4,281 documents were found in the PubMed, Embase, and Scopus databases. While none of the randomised controlled trials of malaria therapies showed an effect on stunting, the evidence from longitudinal observational studies was generally conflicting on the effects of malaria on stunting. In conclusion, there is currently not enough data to incorporate malaria into the Lives Saved Tool (LiST) model as a factor of stunting or to think of malaria therapies as having an impact on stunting.

Ford and Janies (2020) employed ensemble machine learning modeling to predict Plasmodium falciparum isolates exhibiting artemisinin resistance. The study aimed to accurately predict artemisinin drug resistance levels, quantified by IC values, and estimate the parasite clearance rate using in vitro transcriptional profiles. They developed novel machine learning models by transforming isolate data and managing the high-dimensional variables resulting from these transformations. The study demonstrated the effectiveness of ensemble approaches, with a soft voting ensemble model identified as the best-performing model based on its low normalized Root Mean Squared Error (RMSE). Additionally, in vivo transcriptional data from Mok et al. (2015) was utilized to support predictions related to parasite clearance rates. Their findings highlight the

ISSN NUMBER: 1116-249X

possibilities of group learning methods in improving resistance prediction and malaria treatment strategies.

The genotypic profiles of Plasmodium falciparum and Plasmodium vivax isolates collected over a wide geographic area were examined by Rahmasari et al. (2022), along with their relationship to resistance to anti-malarial medications used in Indonesia. Between 1991 and 2022, they conducted a systematic review of the literature. Preferred Reporting Items for Systematic Review and Meta-Analysis, or PRISMA, guidelines were followed in conducting this review, which was documented in accordance with a predetermined methodology. The study's findings indicated that molecular markers of anti-malarial drug resistance were primarily found in P. falciparum and P. vivax. A computerised search was conducted using Pubmed as a reference, combining the terms (falciparum OR vivax resistance, Indonesia) with various combinations of Artemisinin-based combination treatments ACTs (Chloroquine). In order to change chemotherapeutic policy treatment tactics, they suggested giving careful thought to stopping local malaria transmission and dynamic patterns against anti-malarial medications.

Tai and Dhaliwal (2022) proposed predicting individual or population malaria risk using machine learning framework that incorporates genetic mutation locations derived from extensive data about genetic variation. Building on Genome-Wide Association Studies (GWAS), they incorporated single nucleotide polymorphism (SNP) mutation positions to enhance the predictive capacity of genetic risk models. To address the computational challenges posed by large datasets, they implemented a Genetic Algorithm (GA) for efficient hyperparameter tuning. Their study, which utilized the MalariaGEN dataset comprising over 20,000 individuals across 11 populations, demonstrated that including SNP mutation locations significantly improved model accuracy. The results also provided evidence that deep learning models performed comparably to standard

ISSN NUMBER: 1116-249X

machine learning techniques, suggesting viable alternatives for predictive modeling in genetic epidemiology.

Menard and Dondorp (2023) conducted primary research using drug efficacy studies and the Virto sensitivity test to investigate the biological mechanisms underlying artemisinin resistance. Their regression analyses revealed that, due to the considerably shorter plasma half-life of artemisinins (approximately 1 hour) compared with the companion medications in combination treatments (ACTs) based on artemisinin, the decreased sensitivity to artemisinin exposes companion medications to a greater parasite biomass after the typical three day ACT treatment. Consequently, artemisinin resistance contributes to the emergence of resistance against partner medications (Dondorp et al., 2010). Based on these findings, the authors recommended the development and availability of new antimalarial drugs.

Egwu et al. (2023) evaluated antimalarial treatment failure in Ebonyi State, Nigeria, aiming to determine the extent of drug treatment failure in the region. The study employed both survey and in vitro approaches. Structured questionnaires were administered to collect information on participants' drug usage patterns, and a systematic sampling method was applied. The survey captured qualitative data from malaria patients as well as from pharmacies supplying antimalarial drugs. These findings were complemented by in vitro assays assessing the active pharmaceutical ingredient (API) content of artemether/lumefantrine, a popular combination treatment based on artemisinin. The research found that ACTs remain the main therapy for malaria in Nigeria, while non-artemisinin medications are also used. Additionally, it was observed that numerous patients continue to manage symptoms without professional oversight, though this conduct might not be directly linked to the failure of treatment. The authors concluded that regular monitoring of the APIs and the amount of excipients may be useful in understanding the fundamental reasons of

ISSN NUMBER: 1116-249X

failure of treatment and recommended a more thorough survey and drug analysis to completely comprehend Nigerian trends.

Seyoum et al. (2023) used a multilevel binary logistic regression to study the use of insecticide-treated nets (ITNs). Malaria is still a major cause of illness and mortality among children under five in low- and middle-income countries, and ITN use is still low in East Africa, despite the fact that it is avoidable and ITNs are widely accessible. The purpose of this study was to evaluate ITN usage and its contributing factors among local homes with children under five. 174,411 weighted samples from the most recent Demographic and Health Survey (DHS) datasets from East African nations were used in the analysis. A multilevel binary logistic regression model was fitted in light of the data's hierarchical structure, and the model with the lowest deviation was chosen as the best match. The multivariable analysis includes variables that had a p-value of less than 0.2 in the bivariable analysis. The intensity and significance of relationships were reported using adjusted odds ratios (AORs) with 95% confidence intervals (CIs). ITN use ranged from 11.8% in Zimbabwe to 70.03% in Rwanda, with 46.32% (95% CI: 46.08%, 46.55%) of homes with children under five using them, according to the study. Carers between the ages of 25 and 34, marital status (married, widowed, or divorced), primary and higher education levels, wealthier families, smaller household sizes, more children under five, media exposure, and male-headed households were all linked to increased likelihood of ITN use. At the local level, increased ITN use was associated with living in a rural area, having more media exposure, being less impoverished, and having more education. The authors stressed the need to bolster media campaigns encouraging ITN use and suggested giving priority to awareness campaigns for households with no formal education, younger carers, the impoverished, and larger families.

ISSN NUMBER: 1116-249X

3. METHODOLOGY

3.1. Data Collection

This research gathered malaria related information from 500 patients at the Guratopp Rayfield Primary Healthcare Clinic, Jos, Plateau State, Nigeria. The collected data contains features such as ages of the patients (age in numbers), kind of test conducted (Microscopy, MP or malaria Rapid Diagnostic Tests, mRDT), test result (Positive or Negative), choice of malaria drug administered (Artemisinin-based Combination Therapy (ACT or Not_ACT), number of doses administered (dosage numbers) and binary outcome (resistant [1] or not_resistant [0]). At the point of model building, categorical features such as *Test*, *Result* and *Drug* were further expanded and label-encoded to numerical variables using the pandas' built-in *get_dummies* method. These new features subsequently made up the predictors (*age, dosage, test_MP, test_mRDT, result_NEG, result_POS, drug_ACT,* and *drug_Not ACT*) utilized for the model building.

Abbreviations applied in this study are as follows:

NEG = Negative test result

POS = Positive test result

mRDT = Malaria Rapid Diagnostic test

MP = Microscopy test

ACT = Artemisinin-based Combination Therapy

Not ACT = non-ACT antimalarial drugs

ISSN NUMBER: 1116-249X

3.2. Analysis and Model Development

The research examines the application of classification computational learning model in classifying resistance to malaria drugs among patients at the Guratopp Rayfield Primary Healthcare Clinic. The analysis began with using essential Python libraries as pandas, matplotlib, seaborn and scikit-learn into the python environment (VSCode). Afterwards, data inspection and cleaning were conducted through explorative analysis; descriptive statistics, looking out for missing values, duplicates, and understanding the various types and shapes of the data. The data was relatively clean, as there were no missing values and just few outliers were observed on the *age* column (which was trimmed off). Next was analysis and visualization using histogram to visualize the patient's age distribution, bar chart to compare quantities among the categorical variables, and correlation heatmap to visualize relationship between numerical variables. Next was the model building phase which began with converting all categorical variables into numerical using the pandas' *pd.get_dummies* method. Afterwards, the target class (resistance) was inspected and was found to be imbalanced at 69:31. Class imbalance was further addressed by upsampling the minority class using the *resample* method from the *sklearn* library. Data were further partitioned with 80% udes for training and testing purposes with a random state of 11 adopted through the project. Logistic Regression (logit) was first fitted in classifying patients as being resistant to malaria treatment or not, given predictors. With the classification report, performance was assessed using F1-Score, Recall, Accuracy, Recall and Precision. Next was running cross-validation using GridSearchCV to explore the best parameters that resulted in the most reliable classification while minimizing overfitting or underfitting. Finally, to improve reliability and minimize the possibility of overfitting, hyperparameter optimization of the model was performed using GridSearchCV, GridSearchCV is a systematic cross-validation approach that considers all

ISSN NUMBER: 1116-249X

potential combinations of specified parameters, selecting the configuration that yields the best model performance. Confirming that the final model was accurate and generalizable to unseen data.

3.3. Classification Models

3.3.1. Logistic Regression

Logit regression commonly known as Logistic regression is a prevalent procedure commonly employed for modelling binary processes. The logit regression model also helps to provide answers to enquiries concerning the phenomenon under study, as each predictive variable's coefficient elucidates the related effect regarding this variable affecting the results variable, automatically managing other predicting variables effects. Numerous independent factors, $X_1,\ldots,X_k$ can be examined in relation to the variable being predicted using the logistic regression model Y (Agnieszka et al., 2020). The variable Y is binary and only accepts two values (takes the values: 1 and 0) with the value of 1 meaning that a desired event occurs and value 0 indicating an converse event. Also, the regression analysis process helps determine the key contributing factors, which you can ignore, and how they affect each other.

Beyond its application as a classification tool, it is important to present the mathematical formulation of logistic regression to highlight the statistical inference underlying the analysis.

3.3.1.1. Mathematical Framework of Logistic Regression

For binary classification issues, where the result variable can have two alternative outcomes, logistic regression is a frequently used statistical technique. The research, the dependent variable Y represents drug resistance, with Y=1 denoting resistant and Y=0 denoting non-resistant.

ISSN NUMBER: 1116-249X

The probability that a patient exhibits resistance, given a set of predictors, is modeled as

$$P(Y = 1 \mid X) = \pi(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}}$$

Equivalently, the logit transformation is expressed as:

$$\log\left(\frac{\pi(X)}{1 - \pi(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

where:

$\pi(X)$ = probability of resistance

β_0 = intercept

β_1 = coefficient of predictor variable X_1

The parameters $\beta_0, \beta_{1,\ldots,}\beta_k$ were calculated using the Maximum Likelihood Estimation (MLE) technique, which selects the values of β that increases the probability of observing the sample data.

Interpretation of coefficients:

• A positive coefficient ($\beta_1 > 0$) indicates that as predictor X_1 increases, the odds of drug resistance also increase.

• A negative coefficient ($\beta_1 < 0$) implies a decline in the odds of resistance.

Predictors included:

• X_1: patient age

ISSN NUMBER: 1116-249X

• X_2: drug dosage administered

• X_3: type of diagnostic test (mRDT or Microscopy)

• X_4: test result (Positive/Negative

• X_5: drug type (ACT or Not ACT)

3.3.2. DecisionTree Classifier

Decision Tree Learning is a versatile tool used for predictive modeling that is commonly used in supervised learning across different domains, including machine learning, image analysis, and pattern recognition (Vasiliki et al, 2021). Decision trees are typically generated through algorithmic process that identifies optimal splits in the data, based on various conditions. Tree-based Decision Learning constitutes a non-parametric supervised learning model appropriate for classification as well as regression, utilizing decision rules structured as 'if-then-else' statements.

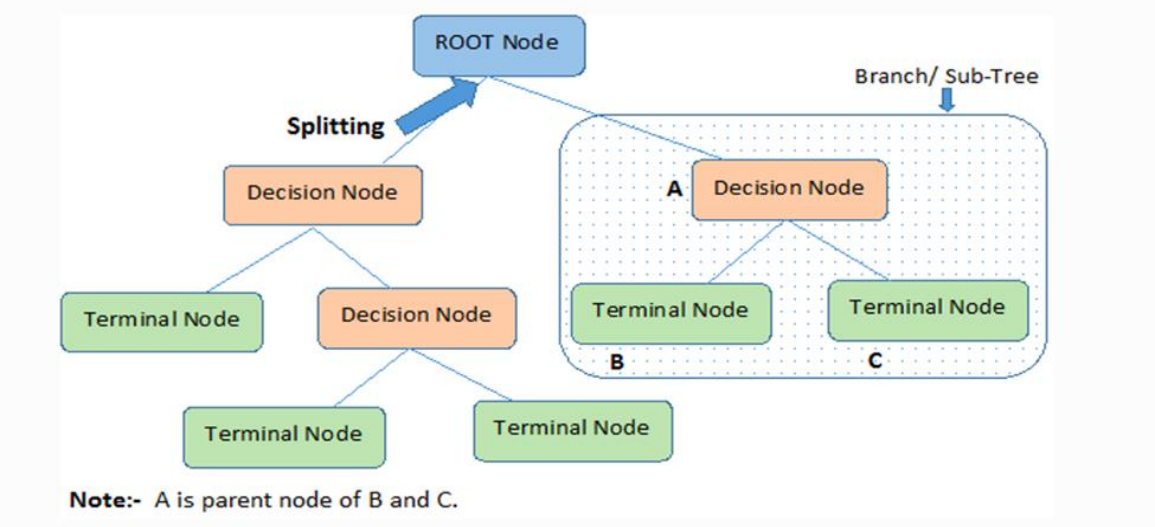

Figure 1: Illustration of the Decision Tree based model

ISSN NUMBER: 1116-249X

Tree-based decision models represents a successive model that unites a set of fundamental assessments efficiently and cohesively where numerical attribute is assessed against a threshold value during each evaluation. The nodes and branches are composed of various tree. Every node signifies attributes within a specific category for classification and every subset establishes a value that could be acquired by the node. Due to their straightforward analysis and accuracy on diverse dataset in various formats, decision trees possess identified numerous areas of execution (Bahzad et al, 2021).

The root node represents the initial split contained in the dataset set. A decision node represents an internal split using a predictor variable. A terminal node(or leaf node) represents the final classification outcome, such as resistant or non-resistant.

3.4. Assessment of the Model and Validation

The process of the design building and validation, involved dividing the dataset into training and validation sets at 80:20 with the random state set at 11 to ensure consistency and reproducibility. The train set comprised 80% (400 observations) of the entire feature data set as well as the independent variables (Predictors) and the resistance class (Response). With 20% of the remaining data (100 observations), reserved for testing the predictive ability of our model on unseen data. This 80:20 split was done to ensure there is a relatively sufficient amount of information during the training and evaluation of the model. The train data were fitted on Logistic Regression (Logit), DecisionTree (DT) classifier, RandomForest (RF) classifier and ensemble techniques like Voting and Staking, as well as GridSearch cross-validation with predictions evaluated against the metrics of Accuracy, Precision, Recall and F1 Score. Behind fitting the classifiers on the train set using the classification metrics shows the Logit model, as having the most accuracy in classifying

ISSN NUMBER: 1116-249X

patients as exhibiting resistance to malaria drugs or not, when tried on new or unseen data. Additionally, cross validation was performed by further partitioning the dataset into several training and validation sets. This gives us a more dependable assessment of how the model performs than a single train-validation split. Since it minimizes the chances of overfitting to a specific group of data, since Logit had the highest accuracy score, GridSearchCV was further fitted on Logit and RF as the base estimators for improved classification. Also, GridSearchCV was particularly selected because we had a smaller set of hyper-parameters to tune and enough computational resources. This cross-validation methodically evaluates all potential combinations of parameters specified, ensuring no potentially optimal combinations was omitted.

3.5. Device Specifications

The analysis and model building were conducted in a Jupyter environment using Microsoft VS Code version 1.94.0 on an installed Python version 3.12.9. Python libraries imported and used pandas for tabular data, Sklearn for machine learning modeling, matplotlib and seaborn for visualization, and numpy for mathematical calculations.

4. RESULT AND DISCUSSION

This section explores the interpretation and discussion of findings from our research. This analysis, visualization and model building was conducted predominantly using Python's built libraries and classes as available on VSCode and Jupyter Notebook's Integrated Development Environment (IDE).

4.1. Result and Interpretation

This result of analysis and model building of this research work are as presented below

ISSN NUMBER: 1116-249X

	age	test	result	drug	dosage	resistance
0	29	mRDT	NEG	Not ACT	0	0
1	21	mRDT	NEG	Not ACT	0	0
2	42	mRDT	POS	ACT	4	0
3	21	mRDT	NEG	Not ACT	0	0
4	60	mRDT	NEG	Not ACT	0	0

Table 1: Data overview

Here, we present an overview (first 5 rows) of the dataset including its corresponding columns ranging from Age, Test, Result, Drug, Dosage and Resistance. Descriptively, average age among the people was 30 years, with the youngest being a year old and the oldest being 95 years old. Also, average doses administered was 2 doses, with few instances of no doses and maximum instances being 4 doses.

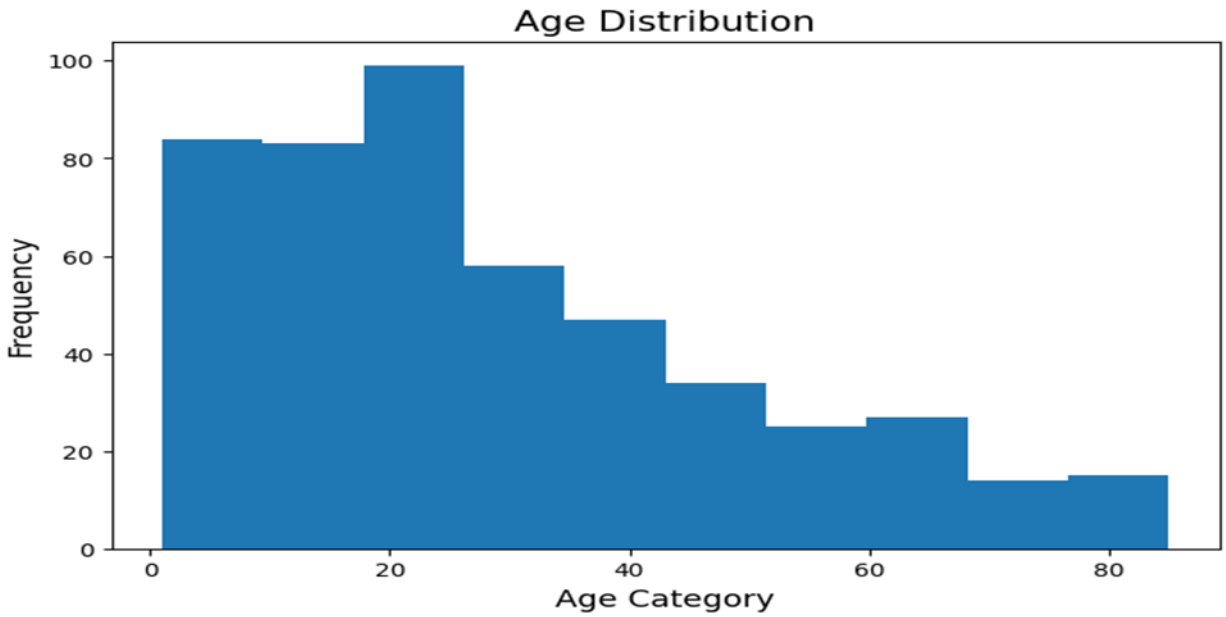

Figure 2: Age Distribution

ISSN NUMBER: 1116-249X

From the histogram distribution above, it could be seen that a vast majority of the patients fall between the age bracket of 10 to 35 years with an average age of about 30 years. Although we had little children (explicitly from age 1) captured in this study, the presence of very few older persons above 85 years presents an outlier that has to be trimmed off. Also, there were more mRDT test (420) conducted than MP tests (80), among the patients. Evaluation of the test result showed that an overwhelming number of the tests came out to be Positive for malaria (306) and Negative (194). Administered drugs similarly revealed a higher number of of ACT malaria drugs administered (306), compared to non_ACT (194). Number of dosages administered also shows that 4 doses are the most frequently administered doses (202), followed by 0 doses (194), 3 doses (49), 2 doses (45) and 1 dose (10). (15) patients showed resistance, while 342 did not. Similarly, Resistance by administered drugs showed that there seem to be more resistance among patients administered ACT based drugs compared to those administered non-ACT based drugs. Resistance was also observed to be considerably more evident in patients who tested positive for malaria than in those who tested negative.

ISSN NUMBER: 1116-249X

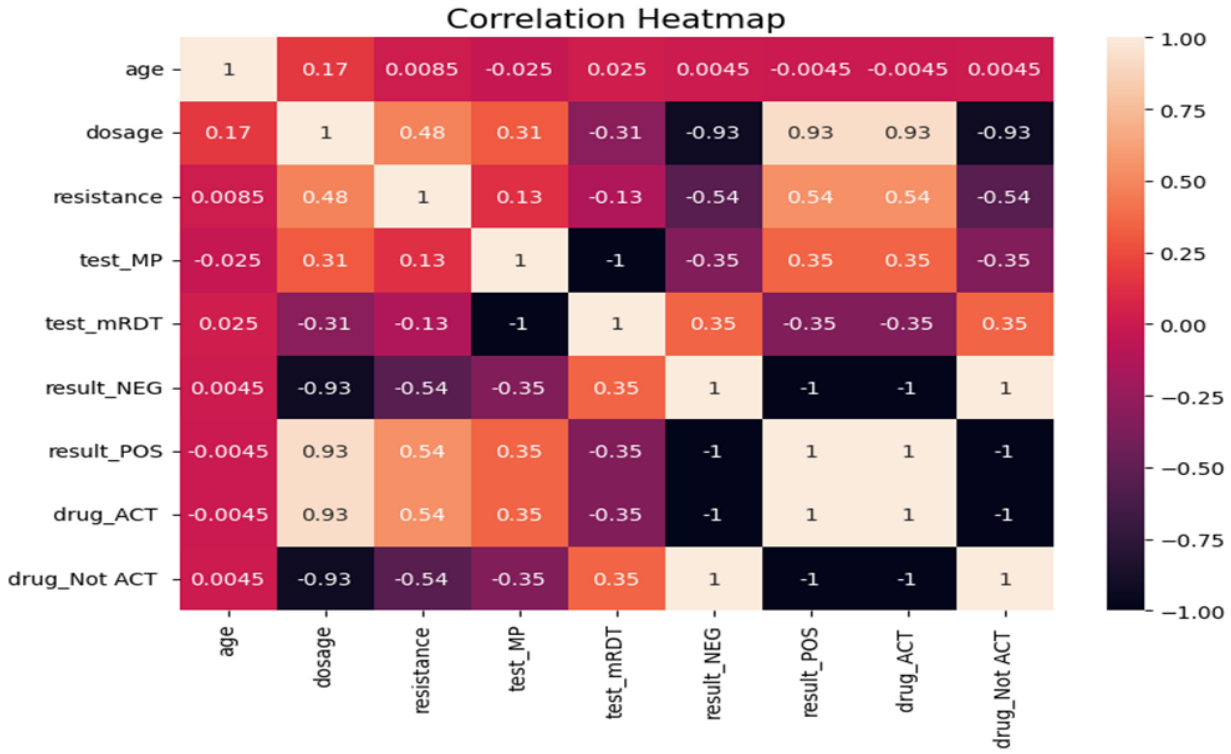

Figure 3: Correlation Heatmap of variables

After label encoding categorical features on the dataset, Figure 3 above illustrates the correlation between the various features and target in our dataset. As seen, administered doses show a moderate positive correlation with resistance (0.48) – meaning the greater the dosage taken, the chances of the patient to becomes resistant to malaria treatment among other factors is higher.

ISSN NUMBER: 1116-249X

```
1 classification(logreg)
✓ 0.1s
Model test Accuracy: 0.74

              precision    recall  f1-score   support

           0       1.00      0.61      0.75        66
           1       0.57      1.00      0.72        34

    accuracy                           0.74       100
   macro avg       0.78      0.80      0.74       100
weighted avg       0.85      0.74      0.74       100
```

Figure 4: Classification Report

The Classification Report as determined by Logistic Regression used in classifying Resistance or Non-Resistance among patients from this study, produced an accuracy score of 74% with higher tendencies in correctly predicting patients non-resistance to malaria as indicated on the Precision and Recall scores of 1.0 and 0.61 respectively.

ISSN NUMBER: 1116-249X

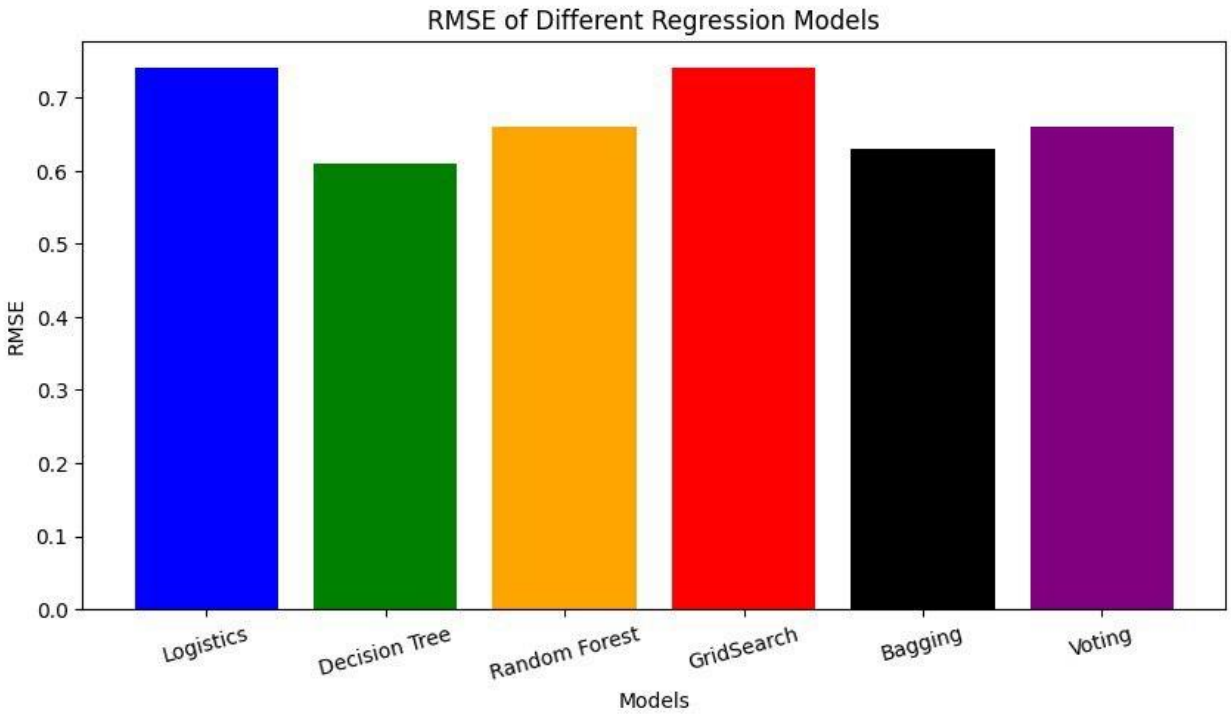

Figure 5: Models and respective Accuracy Scores

Figure 5 above shows the various models built in for predicting the binary resistant and non-resistant classes among patients from our study group. The Logit model had the highest accuracy score at 74% accuracy, in predicting the likelihood of the patients being resistant to malaria drugs or not. Similarly, GridSearch cross-validation estimated on Logit and RF as base estimators, also produced an impressive accuracy score of 74%. The ensembling techniques (Voting and RF) showed 66% accuracy, while Bagging and DT classifiers recorded 63% and 61% respectively. Overall, the GridSearchCV was adopted for predicting the resistance class. It minimizes chances of overfitting and underfitting by fitting multiple versions or subsets of the train set, in arriving at an optimal estimator. Also, because we had a smaller set of hyper-parameters to tune, and enough computational resources.

ISSN NUMBER: 1116-249X

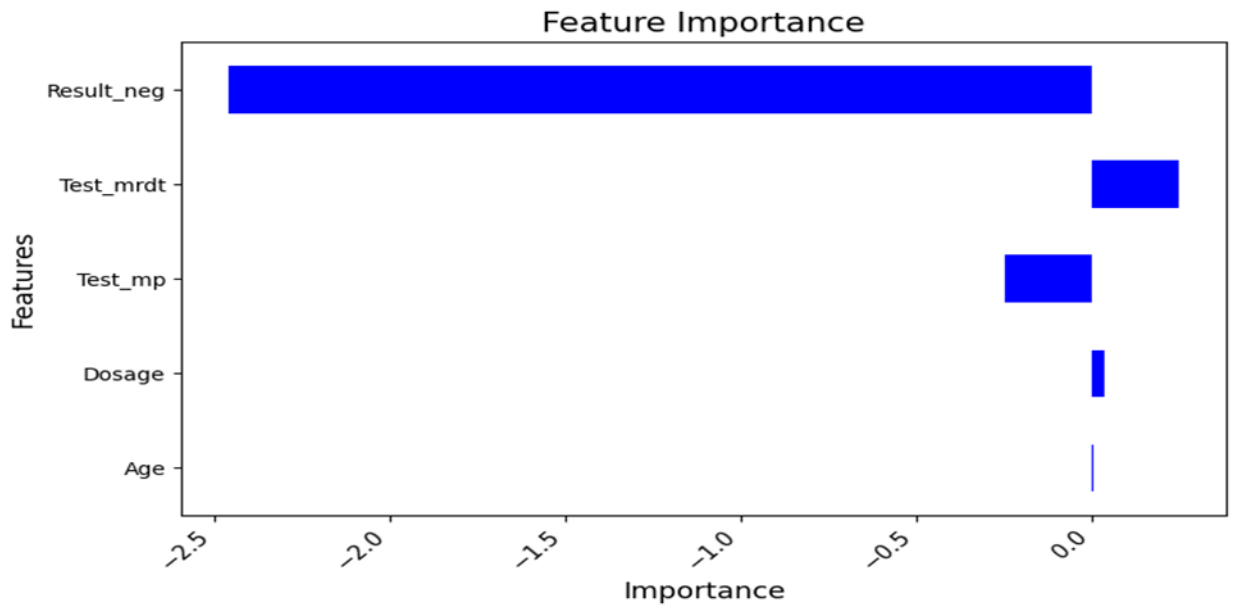

Figure 6: Feature Importances

The Feature Importances figure above, indicates the features (predictors) predominantly used by the Logit model (coefficients) in classifying the patients. As seen, Test Result (negative) is a major factor responsible for predicting patients class. Others, were the choice of test (mRDT and MP), along with test result and age being the set of the factors with minimal influence on predicting patients class, as either being resistant or non-resistant to malaria treatment.

5. SUMMARY, RECOMMENDATION AND CONCLUSION

5.1.Summary

This study investigates malaria trends and treatment outcomes among patients at the Guratopp Rayfield Primary Healthcare Center (PHC) based on data analytics and machine learning. Findings revealed that majority of the patients between 15 to 45 years, with an average age of 28 years and few outlier ages above 85 years, which weren't considered in the analysis. The study reveals the

ISSN NUMBER: 1116-249X

predominant age groups affected by malaria are teenagers and young adults. Similarly, Rapid Diagnostic Tests (mRDT) are markedly more common than Microscopy (MP) tests, with majority of the tests yielding positive results. Artemisinin-based Combination Therapy (ACT) drugs were widely administered, with four doses being the most frequent regimen, thus correlating with higher resistance levels. The study found that patients between 17 and 45 years old exhibited higher resistance to malaria. Those treated with ACT drugs and positive test outcomes were correlated to increased drug resistance while test result and choice of test, also identified as key predictors of malaria resistance among the patients.

5.2. Recommended Approaches

The research proposes that health organizations, clinics, pharmaceutical companies and research institutes should intensify monitoring of ACT resistance and promote proper malaria testing before treatment. Health promotion campaigns should further emphasize adherence to treatment protocols to reduce the probability of resistance development.

In addition, future research should incorporate statistical validation approaches, such as simulation studies, to test the robustness of predictive models across different populations and settings. This will ensure that the findings are not only locally relevant but also generalizable.

5.3. Conclusion

The research revealed that supervised learning models, particularly logistic regression and decision trees, can be applied effectively to assess malaria drug resistance. Findings indicated that age, dosage, test type, and test results were significant predictors of ACT resistance, providing useful insights for healthcare planning and malaria control strategies.

ISSN NUMBER: 1116-249X

Beyond its medical contributions, this work also provides a statistical contribution by showing how logistic regression with parameter estimation and cross-validation can be used to model binary outcomes. The methodology combines predictive accuracy with interpretability, making it adaptable to a wide range of applications beyond malaria research. In this way, the study contributes both to public health and to the advancement of statistical practice.

6. Contribution to knowledge

The study plays a part in the understanding of malaria dynamics in the local context by highlighting the link between patient characteristics and treatment, drug resistance patterns and utilization of machine learning in elucidating these findings. The identification of age and test results as significant predictors of drug resistance supplies useful information for healthcare workers and policymakers to adapt their strategies accordingly. Additionally, the findings on ACT Resistance emphasizes the importance of continuous observation and the possible exploration of alternative treatments. By presenting a comprehensive analysis of malaria data from a healthcare facility, this study extends the current knowledge base in the field of malaria control and management, informing evidence-based decision-making and future research endeavors.

Suggestion for further studies

In other to expand further researches on malaria and malaria drug resistance, quality control tests of pharmaceuticals should be supervised. Consistent evaluation of the effectiveness of antimalarial medications is necessary to guide treatment protocols and to facilitate the prompt identification and reaction to drug resistance. Enhanced monitoring of malaria cases is essential. and deaths to help examine the areas most affected. Application of other machine learning techniques together with cross validations should be explored towards ensuring accurate classifications. Self-

ISSN NUMBER: 1116-249X

medication without tests or hospital and clinical checkup should be avoided. Also, there is need to investigate Sophisticated machine learning models such as deep learning and neural networks have the potential to enhance classification accuracy. By tackling the problem of class imbalance through methods like oversampling, undersampling, or the generation of synthetic data (SMOTE), this could result in a more balanced predictive model. Additional patient information, encompassing medical history, genetic predisposition, and environmental factors, could provide expanded understanding of resistance patterns. Furthermore, long term studies monitoring patient responses over time may facilitate understanding of casual relationships between drug administration and resistance. Lastly, integrating real-world clinical trial data and implementing deep learning techniques could further improve malaria resistance predictions and treatment strategies.

References

1. Agnieszka, S., Agnieszka K. and Danuta, Z. (2020). "Application of logistic regression models to assess household financial decisions regarding debt" Procedia Computer Science, Volume 176, P. 3418-3427, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.09.055.
2. Bahzad, J. and Abdulazeez, A. (2021). "Classification Based on Decision Tree Algorithm for Machine Learning". Journal of Applied Science and Technology Trends. 2. 20-28.
3. Bhatt, S., Weiss, D. J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K. E., Moyes, C. L, Henry, A. and Eckhoff, P. A. (2015). "The Effect of Malaria Control on Plasmodium Falciparum in Africa between 2000 and 2015". Nature 526: 207–211.
4. Chowdary, D. H. (2020). "Decision Trees Explained with a Practical Example". Available at https://pub.towardsai.net/decision-trees-explained-with-a-practical-example-fe47872d3b53 [Last accessed 28/04/2024]
5. Chinedu, O.E., Chinyere, A., Jennifer, C., Joshua, C. N. (2023). "Assessment of the Antimalarial Treatment Failure in Ebonyi State, Southeast Nigeria"

ISSN NUMBER: 1116-249X

6. Greenwood, B. (2010). "Anti-malarial Drugs and the Prevention of Malaria in the Population of Malaria Endemic Areas". Malar J 9 (Suppl 3): S2.
7. Jiang, H., Li, N., Gopalan, V., Zilversmit, M. M., Varma, S., Nagarajan, V., Li, J., Mu, J., Hayton, K., Henschen, B. (2011). "High Recombination Rates and Hotspots in a Plasmodium Falciparum Genetic Cross". Genome Biol 12: pR33. Malaria Drug Resistance
8. Lopera-Mesa, T. M., Doumbia, S., Chiang, S., Zeituni, A. E., Konate, D. S., Doumbouya, M., Keita, A. S., Stepniewska, K., Traore, K., Diakite, S. A. (2013). "Plasmodium Falciparum Clear-Ance Rates in Response to Artesunate in Malian Children with Malaria: Effect of Acquired Immunity". Journal of Infectious Diseases 207: 1655–1663.
9. Miotto, O., Almagro-Garcia, J., Manske, M., Macinnis, B., Campino, S., Rockett, K. A., Amaratunga, C., Lim, P., Suon, S., Sreng, S. (2013). "Multiple Populations of Artemisinin-Resistant Plasmodium Falciparum in Cambodia". Nat Genet 45:648–655.
10. Newton, P. N., Green, M. D., Fernández, F. M. (2010). "Impact of Poor-Quality Medicines in the 'Developing' World". Trends Pharmacol. Sci. 2010, 31, 99–101. [CrossRef]
11. Nosten, F., White, N. J. (2007). "Artemisinin-based Combination Treatment of Falciparum Malaria". Am J Trop Med Hyg 77:181–192.
12. Sarda, V., Kaslow, D. C., Williamson, K. C. (2009). "Approaches to Malaria Vaccine Development using the Retrospectrocope". Infect Immun 77: 3130–3140.
13. Takala-Harrison, S., Laufer, M. K. (2015). "Antimalarial Drug Resistance in Africa: Key Lessons for the Future". Ann NYAcad Sci 1342: 62–67.
14. Thomas, J. A., Tan, M. Y., Bisson, C., Borg, A., Umrekar, T. R., Hackett, F., Hale, V. L., Vizcay-Barrena, G., Fleck, R. A., Snijders, A. P., (2018). "A Protease Cascade Regulates Release of The Human Malaria Parasite Plasmodium Falciparum from Host Red Blood Cells". Nat. Microbiol. 2018, 3, 447–455. [CrossRef] 11.
15. Vasiliki, M. and Efthimios, A. (2021) "Decision tree learning through a Predictive Model for Student Academic Performance in Intelligent M-Learning environments" Computers and Education: Artificial Intelligence, Volume 2, ISSN 2666-920X, https://doi.org/10.1016/j.caeai.2021.100035
16. White, N. J., Pongtavornpinyo, W., Maude, R. J., Saralamba, S., Aguas, R., Stepniewska, K., Lee, S, J., Dondorp, A. M., White, L. J., Day, N. P. (2009). "Hyperparasitaemia and Low Dosing are an Important Source of Anti-Malarial Drug Resistance". Malaria J 8: 253.

ISSN NUMBER: 1116-249X

17. World Health Organization (2023). World Malaria Report 2020: 20 Years of Global Progress and Challenges. Available online: https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2020 [accessed on 30 November 2023]
18. World Health Organization 2022. World Malaria Report
19. Yeung, S., Van Damme, W., Socheat, D., White, N. J., Mills, A. (2008). "Access to Artemisinin Combination Therapy for Malaria in Remote Areas of Cambodia". Malaria J 7: 96.
20. Yewhalaw, D., Wassie, F., Steurbaut, W., Spanoghe, P., Van Bortel, W., Denis, L., Tessema, D.A., Getachew, Y., Coosemans, M., Duchateau, L. "Multiple Insecticide Resistance: An Impediment to Insecticide-Based Malaria Vector Control Program". PLoS ONE 2011, 6, e16066. [CrossRef]

ISSN NUMBER: 1116-249X