Western European Journal of Medicine and Medical Science
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/3
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

MALARIA DISEASE PREDICTION IN WEST AFRICA USING SELECTED MACHINE LEARNING TECHNIQUE

***[1]Onyijen, O. H., [1]Ogieriakhi, O., [1]Awe, Oluwatobi and [2]Olaitan, E.O**
[1]Department of Mathematical and Physical Sciences, Samuel Adegboyega University, Ogwa, Edo State, Nigeria.
[2]Department of Computer Science and Engineering, University of Hull, United Kingdom
Email: greatolaitaneben.oe@gmail.com

***Corresponding Author: ojei.onyijen@gmail.com**

ABSTRACT

Malaria, a life-threatening disease caused by Plasmodium parasite, remains a global health challenge with significant morbidity and modularity, particularly in sub-Saharan Africa. According to estimates from the World Health Organization (WHO), there were approximately 229 million clinical cases of malaria in 2019 and 409,000 deaths as a result (World Malaria Report, 2019). As a result of an increase in cases and fatalities, malaria is becoming a serious public health concern in West Africa. The focus of this study is to ensure machine learning can help people make a preliminary judgement about malaria according to their daily physical examination data and it can serve as a reference for doctors. The dataset was collected from Kaggle public repository and used to develop a predictive supervised machine learning models such as random forest, decision tree, k-nearest neighbor, artificial neural network and gradient boosting algorithms. Gradient boosting and Decision tree models were found to be the best performing model with an accuracy of 98.3% and 91.3% respectively. The evaluation metrics deployed for the study showed that RMSE (0.11), MAE (0.012), MSE (0.012), F1-score(0.80, 1.00). To further strengthen the evaluation method, confusion matrix produced TP 4 ,0, TN 1,76. The model will help health works, medical personnel and even the patients when diagnosing, to correctly predict Malaria among pateients suspected to have malaria.

Keywords: *Malaria, Machine learning, Algorithm, Confusion matrix, Evaluation metrics*

Western European Journal of Medicine and Medical Science
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/3
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

1. Introduction

The expanding field of machine learning is being employed in a variety of medical-related operations. All machine learning models learn from the past and make predictions based on some dataset. Machine learning approaches are commonly employed in malaria prediction and produce favourable outcomes. Malaria detection has become incredibly simple and inexpensive as machine learning advances. There are numerous Malaria-related datasets available and as a result, machine learning is required for use in medical diagnosis. Machine Learning is a branch of computer science that enables computer systems to gradually learn from data without explicit programming. By fusing computer science and information technology, healthcare informatics (HI) works with medical and health data in a contemporary and quickly evolving field. Health care advances in machine learning put it one step ahead of problems (Pollettini *et al.*, 2012).

The four major healthcare applications that can benefit from Machine Learning techniques are prognosis, diagnosis, treatment, and clinical workflow (Ahmad et al., 2018). Prognosis involves predicting the expected development of a disease, the likelihood of survival, identifying symptoms and signs related to a specific disease, and finding out if they will worsen, improve, or remain stable over time (Maity et al., 2017). Diagnosis is a systematic way of identifying a disease by its symptoms and signs. Machine Learning is used in medical treatment to detect the effects of drugs on diseases, creating room for further diagnosis. Then clinical workflow is defined as directed series of steps comprising a clinical process that is performed by people or equipment/computers, and consumes, transforms, and produces information. Some already claim that machine learning and Artificial Intelligence diagnose disease and treat illness earlier and better. Healthcare delivery concerns are most predominant in Africa, and it is imperative that the system of medical diagnosis in Africa must be automated (Araújo et al., 2016). One of the main objectives of Machine Learning in healthcare is to help automate its operation and predict diseases and gain insight from the hidden patterns, facts, or trends that may have been hidden in the data (Skorburg, 2020). Identifying symptoms and signs associated with a particular disease and determining whether they will get worse, get better, or stay the same over time are all parts of the prognosis process. It also includes estimating the likelihood of survival (Maity *et al.*,2017). By studying a disease's symptoms and signs, a diagnosis can be made. In medical therapy, machine learning is used to identify how medications affect illnesses, allowing for more thorough diagnosis.

African countries are particularly affected by issues with healthcare delivery, hence it is essential that the continent's medical diagnosing system be automated. One of the key goals of machine learning in the healthcare industry is to assist in operating it more efficiently, forecast diseases, and gain insight from any hidden patterns, facts, or trends that may have been present in the data (Skorburg, 2020).

Western European Journal of Medicine and Medical Science
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/3
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

Health issues keep increasing rapidly in some countries especially diseases that related to blood disorders. In year 2021, there were 247 million cases of malaria around the world compare to the 245 malaria cases in 2020. The diagnosis of malaria is based on clinical symptoms and/or measurement of plasma glucose. Malaria and its complications can cause severe problems to affected individuals, their families, and they impose a heavy burden on health services. (WHO 2023). According to the National Academies of Sciences and Medicine (2017). Malaria disease continues to be the leading cause of mortality worldwide, particularly in West Africa, hence this study (Abebe *et al.,* 2019). Various studies such as (Adebanji et al., 2021; Suseela et al., 2021; Adamu & Singh, 2021), have deployed machine learning approach to predicting malaria, yet there is a paucity of study in Africa on utilizing machine learning to predict malaria patients. Using a dataset from an online resource Kaggle, a supervised predictive machine learning model based on Random Forest, K-nearest neighbor, Decision Tree, Artificial Neural Network, Gradient Boosting will be deployed for the prediction of malaria using the dataset available.

2.1 Malaria

Malaria is an acute febrile illness caused by Plasmodium parasites, which are spread to people through the bites of infected female Anopheles mosquitoes (WHO, 2023). It is preventable and curable. According to estimates from the World Health Organization (WHO), there were approximately 229 million clinical cases of malaria in 2019 and 409,000 deaths as a result (World Malaria Report, 2019). In many impoverished nations, malaria continues to rank among the deadliest infections that pose a life-threatening threat and has had a significant impact on human life for thousands of years (Molineaux, 1988). Given that the region's estimated incidence and mortality rates are currently around 90% worldwide, Sub-Saharan Africa is a hotspot for its transmission. Plasmodium parasites of four main species *Plasmodium malariae, Plasmodium falciparum, Plasmodium ovale, and Plasmodium vivax* cause malaria, but P. falciparum and P. vivax are the most frequent causes of infection and are primarily transmitted to people through the bite of infected female anopheles' mosquitoes, also known as malaria vectors (White & Ho, 2011). The type of parasite, the vector, the human host, and the environment that supports the life of the vector all have an impact on how intensely malaria is transmitted. Diagnosing malaria requires locating parasites or antigens in the patient's blood. Clinical diagnosis, when the doctor determines the patient's diagnosis based on physical signs and symptoms and other physical exams, is one of the most popular ways to diagnose malaria. That is the most popular and least expensive way. This method, however, presents several difficulties because it lacks specificity in the signs and symptoms of malaria, which could lead to incorrect diagnosis and incorrect therapy (World Malaria Report, 2019). A microscopic diagnostic utilizing Giemsa, Wrights, or Field's staining on thin and thick blood smears is an additional method of diagnosing malaria. It has been in use for more than a century and is known as the "gold standard" for laboratory diagnostics (Tangpukdee *et al.,* 2009).

Western European Journal of Medicine and Medical Science
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/3
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

2.2 Machine Learning Technique

Computer science today has a large focus on machine learning algorithms, and these are demonstrated massively in the field of knowledge Discovery Data (KDD), and generally, it has been the new focus of the WHO to employ technology to improve the healthcare system. The advances in digital technologies are becoming essential tools for healthcare specialists to provide the best care for patients (Toh & Brody, 2021). Drug discovery is the next key research area for the healthcare industry (Mathur, 2018). Research in pharmaceutical companies for certain diseases is continuously growing, and machine learning helps speed up drug discovery by analyzing medicinal data and providing prediction models on drug reactions even before they are injected into subjects in a controlled environment. This saves both time and money, as the simulation of drug reactions gives an estimate on likely cure patterns and reactions to the drug (Mathur, 2018). Machine learning models are rapidly advancing and are useful for predicting and assessing structural performance, identifying the structural condition and informing preemptive and recovery decisions by extracting patterns from data collected via various sources and media. Machine learning tools have become available in diagnosing and predicting diseases, thereby saving costs and improving the likelihood of surviving, especially in some deadly diseases (Elshawi, 2020). In the case of infectious diseases, early diagnosis is highly needed in isolating the subjects to reduce the spread of the disease. Predictive analytics is the next era of application of machine learning in the healthcare industry. The focus would be on predicting the likely number of people who could develop a certain disease at a given time, the age of who may likely develop a certain disease, and find patterns that may indicate a disease's status. This is done by assessing the large volume of datasets and images faster (Waring et al., 2020). Hence, the research on the classification and prediction model of auxiliary diagnosis based on clinical data has become one of the hot spots in intelligent medicine. However, most hospitals are not currently deploying machine learning solutions. One of the reasons for this is that health care professionals often lack the machine learning expertise necessary to build a successful model, deploy it in production, and integrate it with the clinical workflow (Sun et al., 2021). However, employing machine learning in the healthcare sector will help ease the stress on physicians since high volumes make them more error-prone, machines can handle large sets of imaging data with a lower error rate, work in place of doctors in their absence.

According to Liyuan & Jennifer (2018) machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly. In machine learning, algorithms are used to distinguish between meaningful and irrelevant patterns in data.

Western European Journal of Medicine and Medical Science
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/3
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

Examples of machine learning applications include the provision of accurate medical diagnostics (breast cancer), real-time map-based monitoring of environmental disasters (forest fires) and sensory monitoring in the industrial process (mechanical failure). Praveena & Jaiganesh (2017) describe Machine learning as a kind of artificial intelligence (AI) which compose available computers with the efficiency to be trained without being veraciously programmed. ML learning interest on the extensions of computer programs which is capable enough to modify when unprotected to new-fangled data.

2.3 Review of related literature

Several studies were found to have been conducted on the machine learning prediction using data in the health sector. The study of Alabi et al. (2020) on "Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer", the predictive supervised ML approach for the estimation of the risk recurrence in early stages of oral tongue squamous cell carcinoma has been developed. The result of the study showed the ability of supervised ML to predict locoregional recurrences**.** The Study of Onyijen et al. (2023) on data driven machine learning techniques for the prediction of cholera outbreak in West Africa explored the use of machine learning algorithms such as Decision Tree, Random Forest, and Logistics Regression to evaluate the prevalence of cholera epidemics in West Africa countries. The results show that logistic regression has an accuracy of 0.47%, random forest 0.978% and the most efficient model was the decision tree 0.998% with a mean squared error and mean absolute error of 0.001% respectively shows that the model will accurately predict cholera outbreak in Africa. Overall results will improve the understanding of the significant roles of machine learning techniques in healthcare data.

A few studies have also been conducted using machine learning data to predict the prevalence and diagnosis of malaria diseases. The study of Wang et al. (2020) on "A novel model for malaria prediction based on ensemble algorithms" deployed ARIMA, STL+ARIMA, BP-ANN and LSTM network models separately applied in simulations using malaria data and meteorological data in Yunnan Province from 2011 to 2017. They compared the predictive performance of each model through evaluation measures: RMSE, MASE, MAD. In addition, gradient-boosting regression trees (GBRTs) was used to combine the above four models. Their result showed that the root mean square errors (RMSEs) of the four sub-models were 13.176, 14.543, 9.571 and 7.208; the mean absolute scaled errors (MASEs) were 0.469, 0.472, 0.296 and 0.266 and the mean absolute deviation (MAD) were 6.403, 7.658, 5.871 and 5.691. After using the stacking architecture combined with the above four models, the RMSE, MASE and MAD values of the ensemble model decreased to 6.810, 0.224 and 4.625, respectively. The findings suggest that the predictive performance of the final model is superior to that of the other four sub-models, indicating that stacking architecture may have significant implications in infectious disease prediction. Furthermore, the study of Lee et al. (2021) on "Machine

Western European Journal of Medicine and Medical Science
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/3
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

learning model for predicting malaria using clinical information" extracted patient information from the PubMed abstracts from 1956 to 2019. They used two datasets: a solely parasitic disease dataset and total dataset by adding information about other diseases. Six machine learning models were compared such as support vector machine, random forest (RF), multilayered perceptron, AdaBoost, gradient boosting (GB), and CatBoost. In addition, a synthetic minority oversampling technique (SMOTE) was employed to address the data imbalance problem. The outcome revealed that the solely parasitic disease dataset, RF was found to be the best model regardless of using SMOTE. Concerning the total dataset, GB was found to be the best. However, after applying SMOTE, RF performed the best. Considering the imbalanced data, nationality was found to be the most important feature in malaria prediction. In case of the balanced data with SMOTE, the most important feature was symptom. Their results demonstrated that machine learning techniques can be successfully applied to predict malaria using patient information. In the study of

Adebanji et al. (2021) on "A Model For Predicting Malaria Outbreak Using Machine Learning Technique" used five (5) supervised machine learning techniques to model the outbreak of malaria using meteorological and malaria incidence data of collected from 2010 - 2020. The machine learning techniques that was used are, Naive Bayes, Support Vector, Linear Regression, Logistic Regression, and K-Nearest Neighbor. The result of the research shows that Naive Bayes has the best accuracy for both testing and training with average accuracy of 79.1% and therefore is the best prediction model that can be used for predicting malaria incidence outbreak using the data set used in this research, Support Vector machine (SVM) is the second best prediction model that can be used for predicting malaria incidence outbreak for both testing and training data with average accuracy of 75.45%, followed by K-Nearest Neighbor with average accuracy of 70.8%, followed by Logistic Regression prediction model which has an average accuracy of 68%, based on this research work it is not advisable to use Linear Regression prediction model for predicting malaria incidence outbreak because it has an average accuracy of 26.05%. Also, the study of Adamu & Singh (2021) on "Malaria Prediction Model Using Machine Learning Algorithms" focused on weather condition, non-climatic features, and malaria cases are considered in designing the model for prediction purposes and also the performance of six different machine learning classifiers for instance Support Vector Machine, K-Nearest Neighbour, Random Forest, Decision Tree, Logistic Regression, and Naïve Bayes is identified and found that Random Forest is the best with accuracy (97.72%), AUC (98%) AUC, and (100%) precision based on the data set used in the analysis.

The study of Harvey et al. (2021) on "Predicting malaria epidemics in Burkina Faso with machine learning" presented the first data-driven malaria epidemic early warning system that can predict the 13-week case rate in a primary health facility in Burkina Faso. Using the extraordinarily high-fidelity data of infant consultations taken from the Integrated e-Diagnostic Approach (IeDA) system that has been rolled out throughout Burkina Faso, they trained a

Western European Journal of Medicine and Medical Science
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/3
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

combination of Gaussian Processes and Random Forest Regressors to estimate the weekly number of malaria cases over a 13 week period. They found that for the lowest threshold for an epidemic alert, the algorithm has 30% precision with > 99% recall at raising an alert. This rises to > 99% precision and 5% recall for the high alert threshold. Our two-tailed predictions have an average 1σ and 2σ precision of 5 cases and 30 cases respectively.

3.1 Data Pre-processing and Feature Selection

The data collection technique used in this work was Secondary data. Secondary data is any dataset that has been compiled from the Internet, a computer science publication, or relevant statistics data. The dataset was obtained through Kaggle: https://www.kaggle.com/code/aravindvijayaragavan/analysis-of-malaria-data-circa-2000-2017. The training set for the physical examination data consists of eight hundred and fifty-six (856) instances. The information includes five (5) physical examination indexes: the country, the year, number of cases, number of death and the World Health organisation (WHO) region. Python Anaconda software and Jupyter Notebook as editor were used to conduct the analysis. The datasets were initially saved as Microsoft Excel documents. However, they were then pre-processed and prepared in comma-separated values file (CSV) format. Machine learning principles uphold high accuracy through data pre-processing to obtain high quality data (Alexandropoulos et al., 2019). The predicted supervised machine learning models in this work were created using the Python programming language.

3.2 System framework for the dataset using Pythom

This Subsection presents the major frameworks for the proposed system implementation. Predictive supervised ML models for the prediction are Random Forest, Decision Tree, K-Nearest Neighbor, Artificial Neural Networks and Gradient Boosting. The algorithms were directly applied on the dataset with the help of Python programming and its built-in libraries to develop the models. We evaluated the performance of these models using the testing data. Lastly, we performed evaluation metrics to identify the optimal performing models or algorithms.

Western European Journal of Medicine and Medical Science
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/3
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

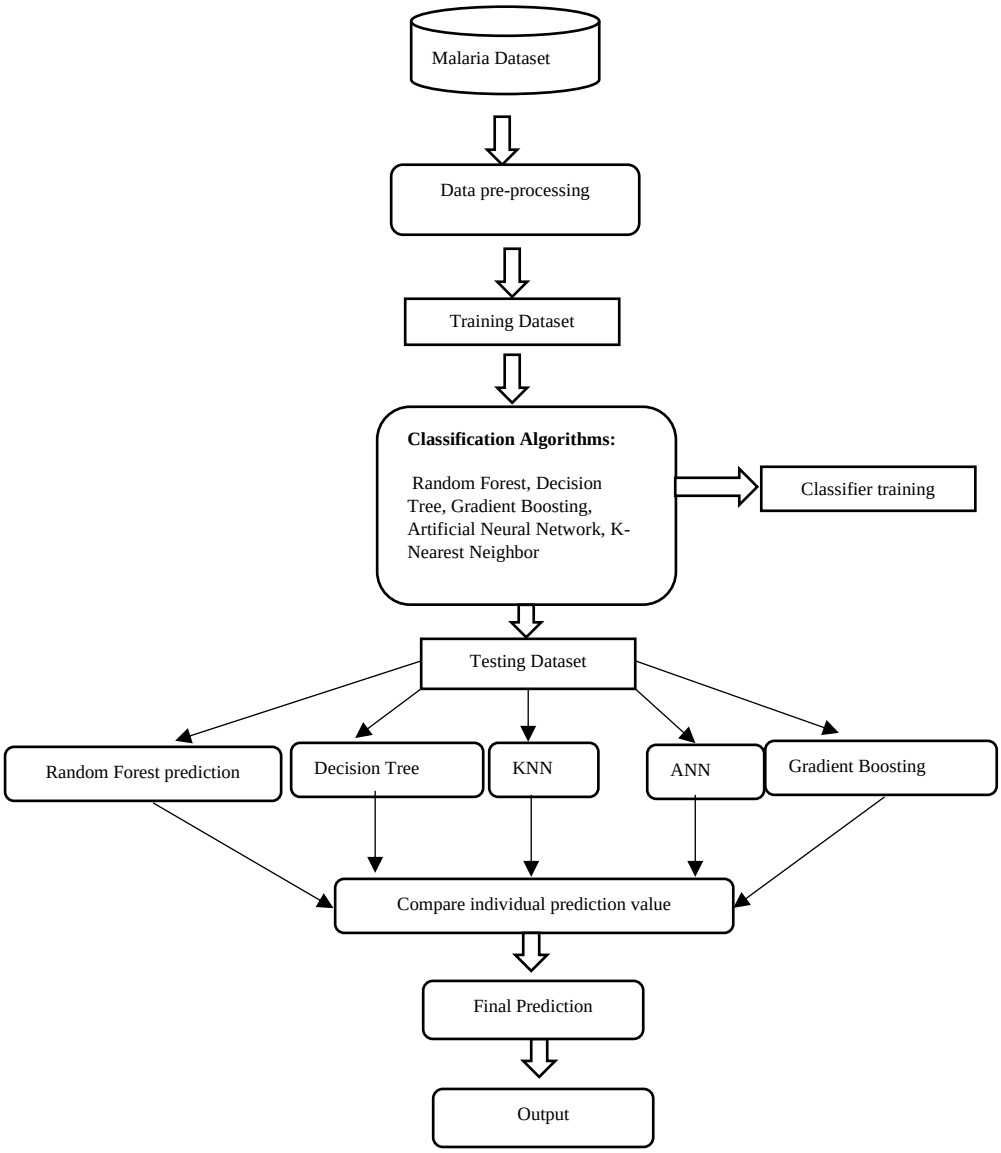


Figure 1 System framework for the dataset

Western European Journal of Medicine and Medical Science
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/3
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

1) Decision Tree

Decision trees is one of the machine learning models. According to Mohamed (2009), decision tree approach or the recursive partitioning algorithm (RPA) is a non- parametric, complex and computerized intensive sorting algorithm. The basic idea is to split the sample responses into the new sub-samples that are as homogeneous as possible and as different from each other, and then to repeatedly split the sub-sample into subgroups until it generates the possibility for decision- making. The entire sample is the root node, while the sub-samples are called nodes. To grow the tree models, greedy algorithm is used such that at each node, evaluate a large set of variable splits so as to find the best split, that is the split that minimizes the weighted decease in impurity:

$$\Delta_i(s,t) = i(t) - P_L i\,(t_L) - P_R i\,(t_R) \qquad \text{Equation 1}$$

Where P_L and P_R denote the proportion observations associated with node t that are sent to the left child node t_L or right child node t_R respectively

2) Random Forest (RF)

Random forests are defined as a group of un-pruned classification or regression trees, trained on bootstrap samples of the training data using random feature selection in the process of tree generation. After a large number of trees have been generated, each tree votes for the most popular class. These tree voting procedures are collectively defined as random forests. A more detailed explanation of how to train a random forest can be found in (Breiman, 2000). For the Random Forests classification technique two parameters require tuning. These are the number of trees and the number of attributes used to grow each tree.

The two meta-parameters that can be set for the Random Forests classification technique are: the number of trees in the forest and the number of attributes (features) used to grow each tree. In the typical construction of a tree, the training set is randomly sampled, then a random number of attributes is chosen with the attribute with the most information gain comprising each node. The tree is then grown until no more nodes can be created due to information loss.

$$nl_j = W_j\ C_j - Wleft_{(j)}C_{left(j)} - W_{right(j)}C_{right(j)} \qquad 2$$

3) K-Nearest Neighbor (KNN)

The algorithm is based on calculation of the number of objects in each class of the sphere (hypersphere) with the centre in the recognized object. The object belongs to the class, which objects dominate in this sphere. This technique supposes that weights have been chosen individually for every object. If weights are not same, instead of calculation of the number of objects their weights can be added together. Thus, if the sphere around the recognized object contains 10 reference objects of class A with the weight 2 and 15 error/boundary objects of Class B with weights 1, the point will be referred to Class A. Weights of objects in the sphere can be expressed as inversely proportional to their distance to the recognized object. Thus, the closer is an object, the more significant it is for this recognized object.

Western European Journal of Medicine and Medical Science
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/3
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

$$d(X,Y) = \{(X-Y)^T\ (I + D_{ww^T}\ (X-Y))1/2\} \quad 3$$

4) Gradient boosting

Gradient boosting is an ensemble algorithm that improves the accuracy of a predictive function through incremental minimisation of the error term (Witten et al., 2011). After the initial base learner (most commonly a tree) is grown, each tree in the series is fit to the so-called "pseudo residuals" of the prediction from the earlier trees with the purpose of reducing the error. The estimated probabilities are adjusted by weight estimates, and the weight estimates are increased when the previous model misclassified a response. This leads to the following model:

$$F(X) = G_0 + \beta_1 T_1\ (X) + \beta_2 T_2\ (X) + \cdots + \beta_k T_k \quad \text{Equation 3}$$

Where G_0 equals the first value for the series T_1 ... T_k, are the trees fitted to the pseudo residual, and β_1 are coefficient for the respective tree nodes computed by the Gradient Boosting algorithm.

5) Artificial Neural Network

Specifically, ANN models simulate the electrical activity of the brain and nervous system. Processing elements known as either a neurode or perceptron are connected to other processing elements. Typically, the neurodes are arranged in a layer or vector, with the output of one layer serving as the input to the next layer and possibly other layers. A neurode may be connected to all or a subset of the neurodes in the subsequent layer, with these connections simulating the synaptic connections of the brain. Weighted data signals entering a neurode simulate the electrical excitation of a nerve cell and consequently the transference of information within the network or brain. The input values to a processing element, i_n, are multiplied by a connection weight, $w_{n,m}$, that simulates the strengthening of neural pathways in the brain. It is through the adjustment of the connection strengths or weights that learning is emulated in ANNs. The basic equation of artificial neural network can be represented as follows:

$$Z = W_x + b_a = f(z)$$

x represents the input vector to the neural network

w represents the weight matrix that contains the weights associated with each connection between the input and hidden layers (or between hidden layers in deeper networks)

b represents the bias vector that is added to the weighted sum.

Z represents the weighted sum of the inputs and biases.

f(z) represents the activation function applied to the weighted sum to introduce non-linearity

A represents the output of the activation function which serves as the input to the next layer.

The above equations describe the computation that occurs at each neuron in a feedforward neural network. The output of one layer serves as the input to the next layer until the final output layer is reached.

Western European Journal of Medicine and Medical Science
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/3
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

3.3 Performance Metrics

Performance metrics are used to ensure that the learning algorithm has learned enough to predict or classify data accurately without minimal error. The performance of the model is evaluated using the following metrics:

1. Precision: It is the number of true positive (TP) results divided by the total actual positive (true positive and false positive).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (4)$$

$$Precision = \frac{True\ Positive}{Total\ Predicted\ Positive} \quad (5)$$

2. Recall is the ratio of correctly predicted outcomes to all predictions. It is also known as sensitivity or specificity.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (6)$$

$$Recall = \frac{True\ Positive}{Total\ Actual\ Positive} \quad (7)$$

3. The F1 score combines these three metrics into one single metric that ranges from 0 to 1 and it considers both Precision and Recall.

$$F1 = 2 * \frac{Precision - Recall}{Precision + Recall} \quad (8)$$

4. Accuracy is a performance metric for a machine classification model defined as the ratio of true positives to true negatives for all positive and negative observations, In other words, accuracy tells us how often we can expect our machine learning model to correctly predict the outcome of the total number of times it made the prediction.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Negative + True\ Negative + False\ Positive} \quad (9)$$

5. Mean Squared Error (MSE) is calculated by taking the average of the square of the difference between the original and predicted values of the data. Mean Absolute Error is the sum of absolute errors over the length of observations/predictions.

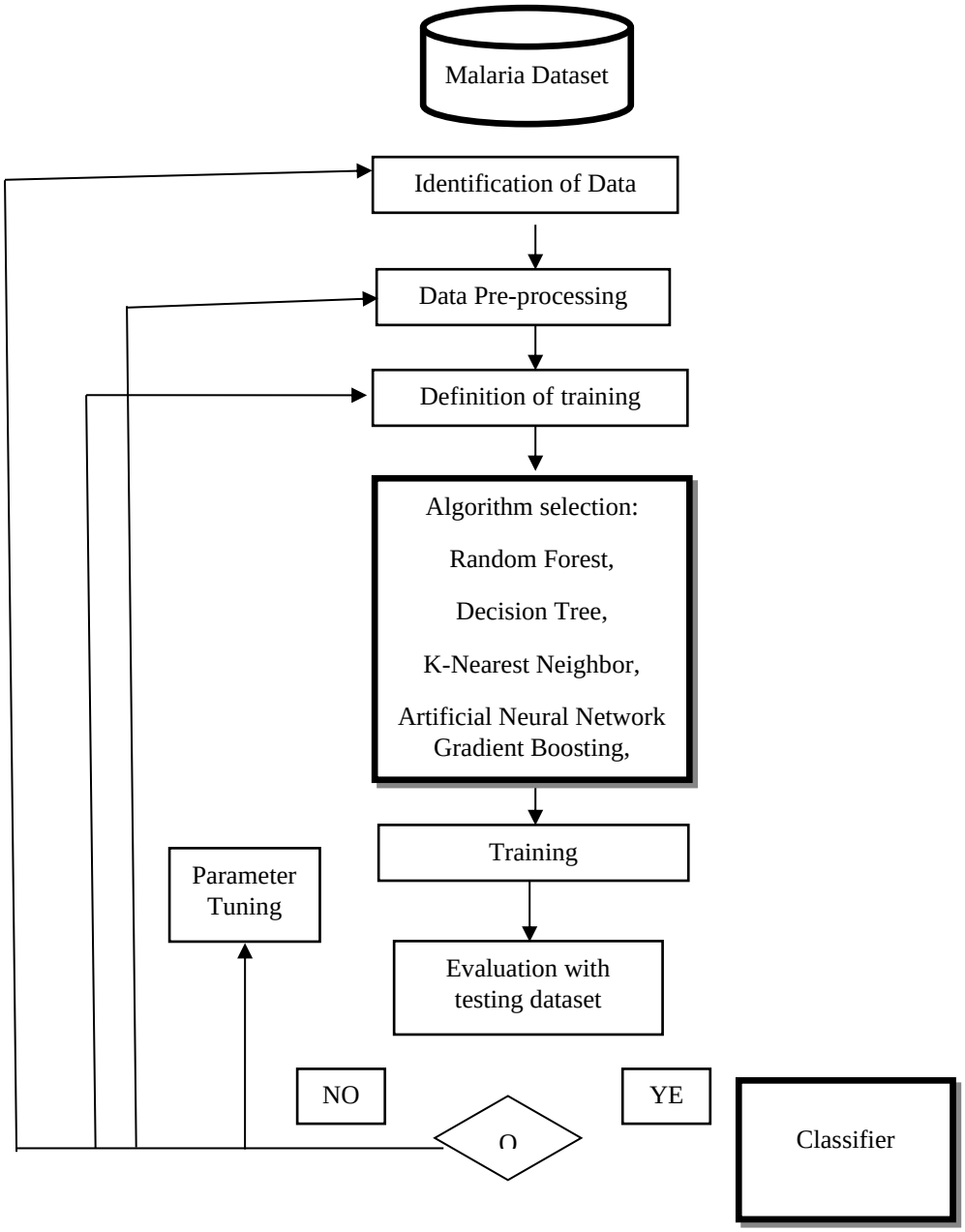
$$\frac{1}{N} \sum_{i=1}^{n} (actual\ values - predicted\ values)^2 \quad (10)$$

Western European Journal of Medicine and Medical Science
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/3
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

3.4 Data visualization

Data visualisation is a technique used to provide quick and effective communication of information in a common way using visual information. This technique helps to identify the hidden relationship amongst variables quicker and highlight areas that need improvement or need more attention. It is used to visualize trends, variabilities and derive meaningful insights from the data, associations, and degree to which each variable affects the other. Visualization of the dataset is done through the integration of libraries such as Seaborn, Plotly, and Matplotlib.

3.5 The process flow of the system



Western European Journal of Medicine and Medical Science
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/3
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

4.0 Results and Discussion

This section covers results and discussion of the data visualisation and the different machine learning model with their performance analysis.

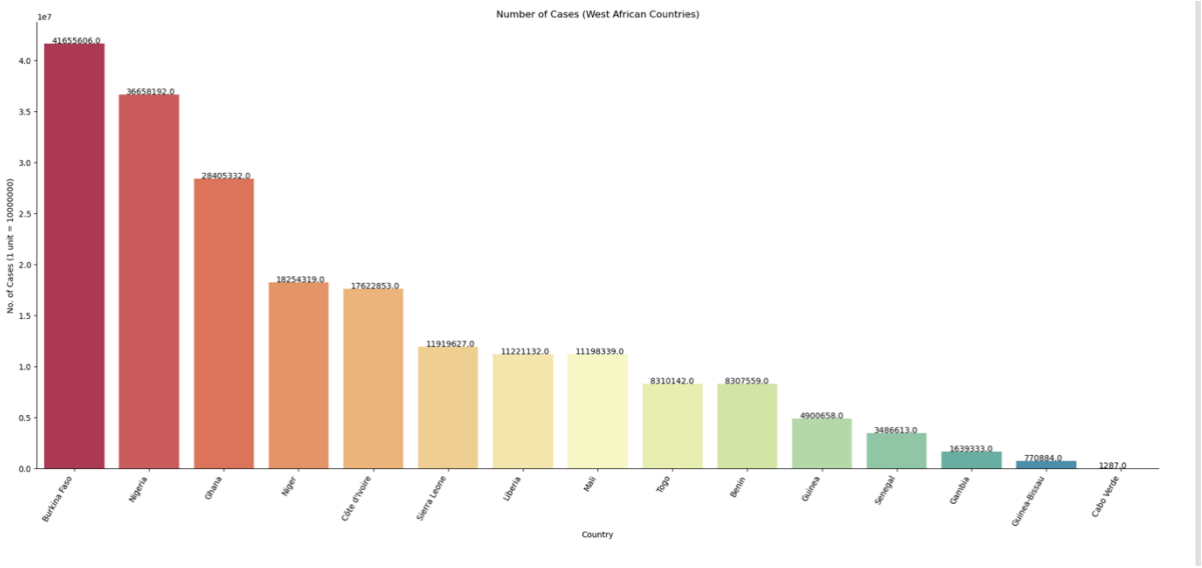


Figure 1: Bar chart of the number of cases and countries attribute

The figure above revealed that Burkina Faso has the highest number of malaria cases in West Africa closely followed by Nigeria. From the result, Cape Verde has the lowest number of cases.

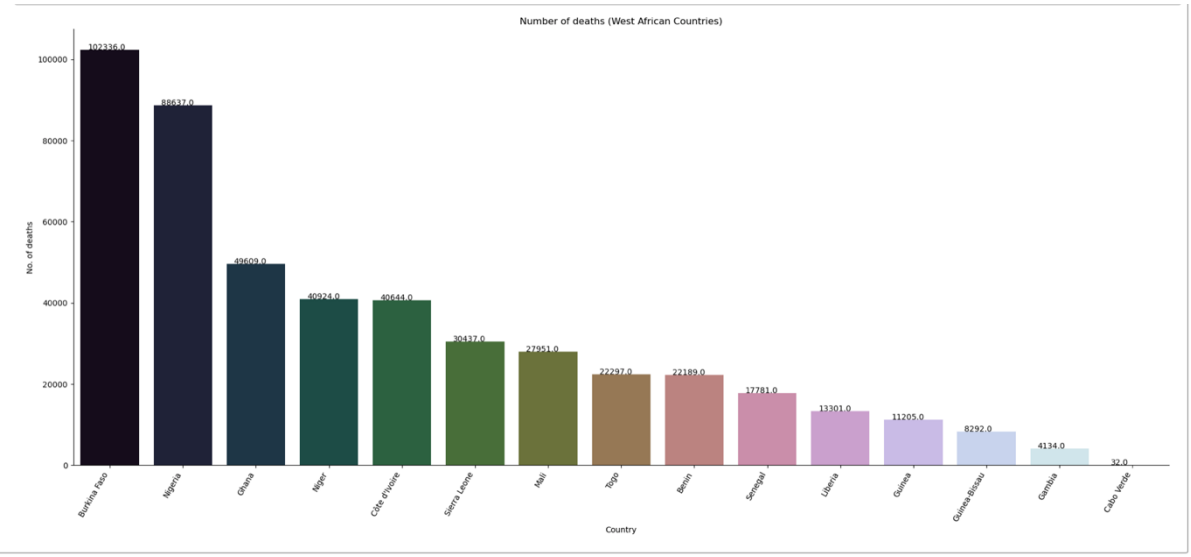


Figure 2: Bar chart of the number of deaths and countries attribute

Western European Journal of Medicine and Medical Science
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/3
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

Figure 2 shows the bar chart of the numbers of deaths as a result of malaria disease in West Africa. It revealed that Burkina Faso also recorded the highest number of deaths as a result of malaria disease. This could be as a result of poor preventive maintenance culture.

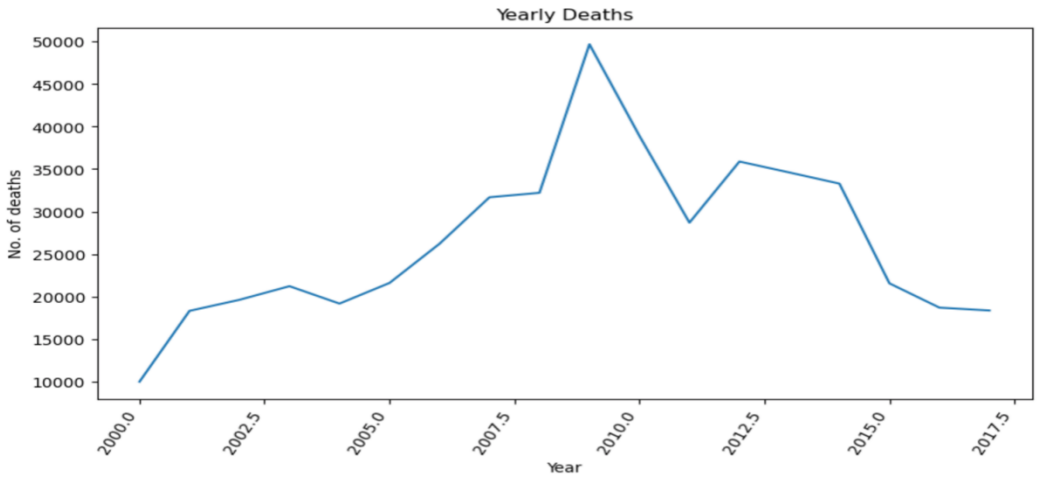


Figure 3: Line graph of the yearly death attribute

The line graph of yearly death attributes for malaria cases between the year of study as shown by Figure 3 revealed that there was a steady increase of malaria death from 2000 to 2007 and peaked between year 2008 and 2009. Furthermore, it showed that a declined was achieved between 2015 and 2017.

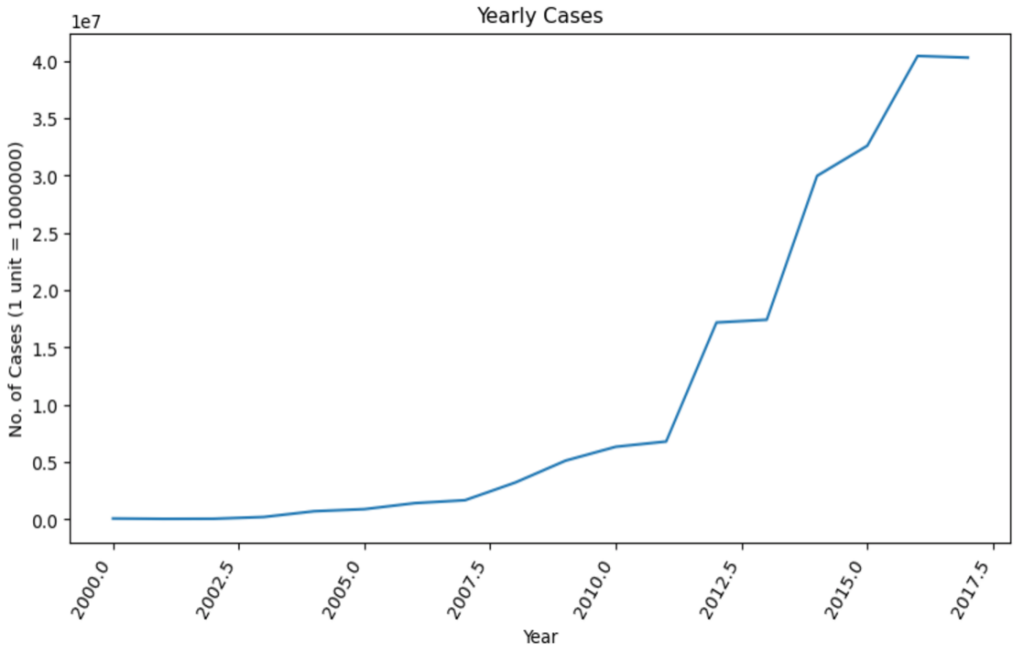


Figure 4: Line graph of the yearly cases attribute

Western European Journal of Medicine and Medical Science
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/3
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

Figure 4 showed the line graph of yearly malaria cases for the study. It revealed that malaria cases increased tremendously between 2011 and 2016 to an estimate of 4 million cases.

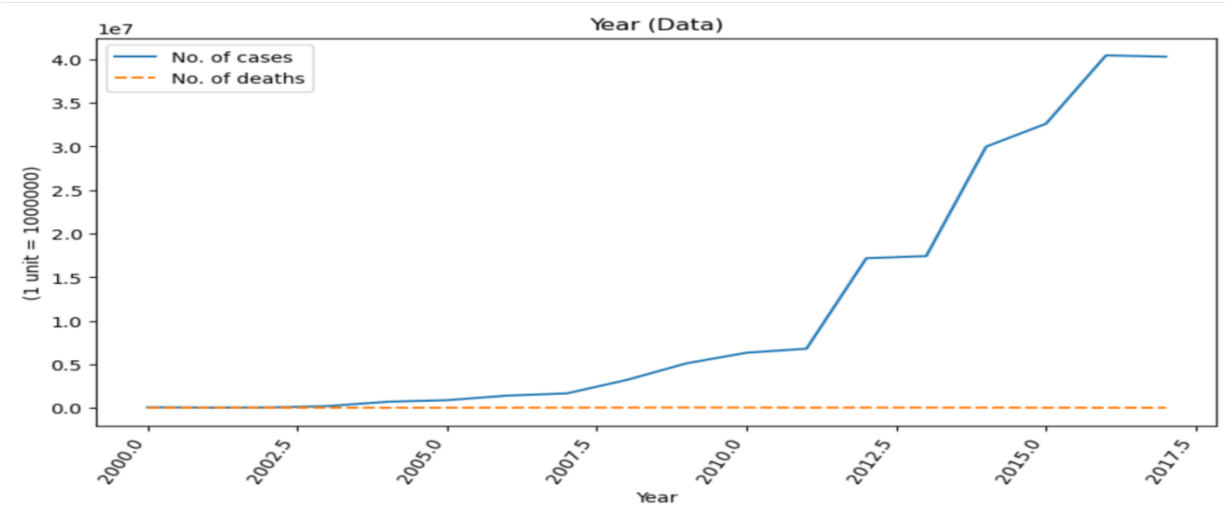


Figure 5: Line graph of the yearly data attribute

Figure 5 captured the comparison of the number of malaria cases and deaths for the period of study. It shows that as malaria cases increases death cases remain fairly constant at a minimal level.

Table 1: Prediction scores for the machine learning models

Models		Precision	Recall	F1-score
Random Forest	0	0.00	0.50	0.12
	1	1.00	0.91	0.99
KNN	0	0.00	0.20	0.08
	1	1.00		0.94
Decision Tree	0	0.00	0.69	0.75
	1	1.00	0.96	0.94
Gradient Boosting	0	0.00	1.00	0.80
	1	1.00	0.99	1.00

Table 2: Performance evaluation result of the model

Metrics	RF	KNN	Decision Tree	ANN	Gradient
Accuracy	90.1	81.5	91.4	3.7	98.8
Mean absolute error	0.099	0.185	0.086	0.00	0.012
Mean squared error	0.099	0.185	0.086	0.00	0.012
Root mean squared error	0.314	0.430	0.294	0.00	0.111

Western European Journal of Medicine and Medical Science
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/3
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

Table 3: Confusion Matrices for Classifier Algorithms

Confusion Matrix for Random Forest

Positive	Negative	**Classified as**
1	7	Positive
1	72	Negative

Confusion Matrix for KNN

Positive	Negative	**Classified as**
1	11	Positive
4	65	Negative

Confusion Matrix for Decision Tree

Positive	Negative	**Classified as**
9	3	Positive
4	65	Negative

Confusion Matrix for Gradient boosting

Positive	Negative	**Classified as**
4	1	Positive
0	76	Negative

Table 3 shows the confusion matrix evaluation metric used for the prediction of malaria cases in West Africa based on the selected machine learning technique.

4.1 Performance Analysis

The study was conducted to predict malaria cases and death rates in West Africa. The data from all malaria outbreaks in each West African country from 2007 to 2017 was gathered to predict the best machine learning model for malaria disease. Previous research had demonstrated the utility of machine learning models in predicting malaria outbreaks globally, but not specifically in West Africa.

The data presented in Figure 1 shows the distribution of malaria cases and death rate in each country. Burkina Faso had a total of 41655606 cases. Figure 2 illustrates the malaria death cases in Africa, with Burkina Faso having the highest death rate of over 100,000. Specifically, it is a measure of the degree to which two variables are linearly correlated.

The dataset was subjected to five machine learning classifier methods: Decision Tree, Random Forest, K-nearest neighbour, Gradient boosting and Artificial neural network to determine the

Western European Journal of Medicine and Medical Science
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/3
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

best performing algorithm for the prediction of malaria disease. Gradient boosting model exhibited the best accuracy of 98.8% and could be utilized to assess the death rate and number of malaria cases. To affirm the Gradient Boosting model's accuracy, the mean absolute error (MAE) of 0.012% and mean square error (MSE) of 0.012% were evaluated and shown in Table 2. The Decision Tree model showed a good accuracy value of 91.4%, with a mean absolute error (MAE) of 0.086% and a mean square error of 0.086% (MSE) in comparison to Random Forest, as depicted in Table 2. Table 1 displays the precision, Recall, F1-score, and support. Precision is defined as the ratio of correctly classified positive samples (True Positive) to a total number of classified positive samples (either correctly or incorrectly). The precision, recall, and F1-score values for the Decision Tree, Random Forest, K-nearest neighbour, Gradient boosting and Artificial neural network models are presented in Table 1. The results indicate that the Gradient Boosting has the highest precision, recall, and F1-score, indicating its superiority over the other models. The support value indicates the number of occurrences of each class in the dataset.

Confusion matrix in Table 3 showed that Random Forest classified 7 malaria positive cases as positive and 1 negative case as positive, while it classified 72 malaria positive cases as negative and 1 negative cases as negative. In the case of k-Nearest Neighbour algorithm, 11 malaria positive cases were classified as positive and 1 negative cases as positive, while it classified 65 positive cases as negative and 4 negative cases as negative. Decision Tree algorithm reveals that 3 positive cases were confirmed positive and 9 negative cases were flagged positive, while 65 positive cases were absolutely negative and 4 negatives cases were flagged negative. Gradient Boosting algorithm reveals that 1 positive cases were confirmed positive and 4 negative cases were flagged positive, while it classified 0 positive cases as negative and 76 negative cases as negative.

In conclusion, the research demonstrates that machine learning models can be effectively utilized to predict malaria in West Africa. Gradient Boosting model exhibited the highest accuracy, precision, recall, and F1-score, making it the most suitable model for malaria prediction. The study's findings provide valuable insights for public health officials and policymakers, enabling them to take proactive measures to prevent malaria in West Africa.

5.0 Conclusion

The data-driven machine learning approaches have been used for the prediction of malaria in West Africa with the hope of reducing the spread of this disease. In predicting the malaria outbreak in West Africa five (5) machine learning models such as Decision Tree, Random Forest, K-nearest neighbour, Gradient boosting and Artificial neural network were used. Grdient Bossting predictive learning-based model was found to be the best model among the developed models with 98.8% in terms of accuracy, mean absolute error 0.012(MSE) and mean squared error 0.012(MSE). This model will help medical personnel predict malaria more accurately.

Western European Journal of Medicine and Medical Science
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/3
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

The model under consideration herein was meticulously designed and comprehensively tested with the aid of a case study dataset. However, it is imperative to note that the accuracy of the proposed model can be further augmented by obtaining additional datasets from a plethora of diverse or multiple case studies. Decision Tree, Random Forest, K-nearest neighbour, Gradient boosting and Artificial neural network Classification Machine learning algorithms were utilized in proposing the model. It is noteworthy that exploring the possibility of using an alternative ML algorithm is highly recommended to ascertain the accuracy rate, and more benchmarking is warranted. It is highly recommended to employ mixed-method research to gain a profound understanding of the case study and also to mitigate the risk of being biased.

REFERENCES

1. Adamu, Y. A., & Singh, J. (2021). Malaria prediction model using advanced ensemble machine learning techniques. Journal of medical pharmaceutical and allied sciences, 10(6), 3794-3801.
2. Adebanji, S., Akomolafe, P. O., & Ogundoyin, K. I. (2021) a model for predicting malaria outbreak using machine learning technique- Retrieved on June 23 2023 fromhttps://anale-informatica.tibiscus.ro/download/lucrari/Vol19/19-1-01Adebanji.pdf.
3. Alabi, R. O., Elmusrati, M., Sawazaki-Calone, I., Kowalski, L. P., Haglund, C., Coletta, R. D., Mäkitie, A. A., Salo, T., Almangush, A., & Leivo, I. (2020). Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer. International Journal of Medical Information, 136, 104068 doi: 10.1016/j.ijmedinf.2019.104068.
4. Alexandropoulos, S. A. N., Kotsiantis, S. B., & Vrahatis, M. N. (2019). Data preprocessing in predictive data mining. *In Knowledge Engineering Review*, 34 (1). https://doi.org/10.1017/S026988891800036X
5. Breiman, L. (2000). Randomizing outputs to increase prediction accuracy. *Machine Learning*, 40:229–242.
6. Harvey, D., Valkenburg, W., & Amara, A. (2021). Predicting malaria epidemics in Burkina Faso with machine learning. *PLoS ONE*, 16(6), e0253302. https://doi.org/10.1371/journal.pone.0253302.
7. Lee, Y. W., Choi, J. W., & Shin, E. H. (2021). Machine learning model for predicting malaria using clinical information. *Computer Biology Medicine*, 129, 104151. doi: 10.1016/j.compbiomed.2020.104151.
8. Liyuan, L., & Jennifer, P.L. (2018). A Comparison of Machine Learning Algorithms for Prediction of Past Due Service in Commercial Credit. Grey Literature from PhD thesis. Retrieved on July 26, 2023 from https://digitalcommons.kennesaw.edu/dataphdgreylit/8.
9. Maity, N.G., & Das, S. (2017). Machine learning for improved diagnosis and prognosis in healthcare. 2017 Institute of Electrical and Electronics Engineers, Aerospace Conference, Held on 4-11 March 2017 in Big Sky, Montana, USA.

Western European Journal of Medicine and Medical Science
Volume 1, Issue 1, September, 2023
https://westerneuropeanstudies.com/index.php/3
ISSN (E): XXXX-XXXX Open Access| Peer Reviewed

10. Mohamed, A. H. (2009). Credit Risk Modeling in Developing Economy: The Case of Libya. PhD Thesis. Griffith University.
11. Molineaux, L. (1988) The Epidemiology of Human Malaria as an Explanation of Its
    a. nearest neighbor algorithm-based cloud-edge computing for cyber-physical-social systems. *Institute of Electrical and Electronics Engineers Access, 8*, 50118-50130.
12. Pollettini, J. T., Panico, S. R., Daneluzzi, J. C., Tinós, R., Baranauskas, J. A., & Macedo, A.
    A. (2012). Using machine learning classifiers to assist healthcare-related decisions: classification of electronic patient records. Journal of medical systems, 36, 3861-3874.
13. Praveena, M., & Jaiganesh, V. (2017). A Literature Review on Supervised Machine Learning Algorithms and Boosting Process. *International Journal of Computer Applications*, 169(8), 0975 – 8887.
14. Skorburg, J. A. (2020). What Counts as "Clinical Data" in Machine Learning Healthcare
    a. Sun, H., Gui, D., Yan, B., Liu, Y., Liao, W., Zhu, Y., & Zhao, N. (2016). Assessing the potential of random forest method for estimating solar radiation using air pollution index. *Energy Conversion and Management*, *119*, 121-129.
15. Suseela, D. S., Samantha, E., Priyadharshini, B., & Jetlin, C. P. (2021). MALARIA DETECTION USING MACHINE LEARNING WITH K-NN ALGORITHM", International Journal of Science & Engineering Development Research, 6(3), 457 - 460.
16. Tangpukdee, N., Duangdee, C., Wilairatana, P., & Krudsood, S. (2009). Malaria diagnosis: a brief review. *The Korean journal of parasitology*, 47(2), 93–102. https://doi.org/10.3347/kjp.2009.47.2.93
17. Wang M, Wang H, Wang J, Liu H, Lu R, Duan T, Gong X, Feng S, Liu Y, Cui Z, Li C, & Ma J. (2019). A novel model for malaria prediction based on ensemble algorithms. *PLoS One*, 14(12):0226910. doi: 10.1371/journal.pone.0226910.
18. Witten, I., Frank, E., Hall, M., & Christopher, J. (2011). *Data Mining: Practical machine learning tools and techniques*, 3rd Edition. Morgan Kaufmann, San Francisco (CA). Retrieved 2011-01-19.
19. World Malaria Report, 2019 *Malaria Report* retrived from https ://www.mmv.org/newsroom/news retrived on May, 2023resources-search/world-malaria-report.
20. Zhang, W., Chen, X., Liu, Y., & Xi, Q. (2020). *A distributed storage and computation k-Distribution, Including Some Implications for Its Control*. In: Wernsdorfer, W.H. and McGregor, I., Eds., Malaria, Principles and Practice of Malariology, Churchill Livingstone, New York.