ANALYSING MACHINE LEARNING BASED MODELS FOR PREDICTING MALARIAL FEVER PRIOR TO CLINICAL TRIAL

BY

Md. Robiul Islam
ID: 191-15-2611

AND
Tanjina Nur Jeba
ID: 191-15-2344

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Md. Sabab Zulfiker
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Co-Supervised By

Dr. S.M. Aminul Haque
Associate professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

February 2023

APPROVAL

This Project/internship titled **"A**nalysing machine learning based models for predicting malarial fever prior to clinical trial**"**, submitted by Md. Robiul Islam, ID No: 191-15-2611 and Tanjina Nur Jeba, ID No: 191-15-2344 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 01 February 2023.

BOARD OF EXAMINERS

4/02/23

Chairman

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Tania Khatun

Tania Khatun (TK)
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

01.02.23

Ms. Lamia Rukhsara (LR)
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Mohammad Shorif Uddin
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

By signing this document, we certify that we completed this project “**A**nalysing Machine Learning based models for predicting malarial fever prior to clinical trial” under the supervision of Md. Sabab Zulfiker, Lecturer (Senior Scale), Department of CSE Daffodil International University. Additionally, we certify that no portion of this project or any element of it has been submitted to another institution for the purpose of receiving a degree or certification.

Supervised by:

04.02.2023

Md. Sabab Zulfiker
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Co-Supervised by:
Dr. S. M. Aminul Haque
Associate Professor & Associate Head
Department of CSE
Daffodil International University

Submitted by:

Robiul islam

Md. Robiul Islam
ID: 191-15-2611
Department of CSE
Daffodil International University

Tanjina Nur Jeba

Tanjina Nur Jeba
ID: 191-15-2344
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

We first want to give God the highest praise for His wonderful gift, which enabled us to successfully finish the final project.

We are really appreciative and like to express our heartfelt gratitude to Md. Sabab Zulfiker, Lecturer (Senior Scale), Department of CSE, Daffodil International University, Dhaka. Our supervisor's deep knowledge and significant interest in the field of "Machine Learning" are required to complete this project. His unending patience, scholarly guidance, constant encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages, and reading many inferior drafts and correcting them at all stages enabled us to complete this project.

We would like to express our heartiest gratitude to Md. Sabab Zulfiker, Dr. S.M. Aminul Haque**,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Malaria is a very dangerous illness brought on by unicellular protozoan parasites of the species Plasmodium. This disease is endemic in many parts of the world. Confirming the presence of parasites early on in all cases of malaria permits the delivery of species-specific antimalarial medication, which reduces the death rate and points to other illnesses in situations where the diagnosis is negative. Nevertheless, light microscopy of thin and thick peripheral blood (PB) films stained with May-Grünwald–Giemsa (MGG) is still the gold standard. Since this is a labor-intensive process that relies on a pathologist's expertise, medical professionals in regions of the world where malaria is not widespread may have difficulties diagnosing cases of the disease. To predict malaria fever, this research used a total of thirteen different machine-learning models. These models included the Gaussian NB, Logistic Regression, XGB, Bagging Classifier, Random Forest Classifier, Extra Trees Classifier, Gradient Boosting Classifier, Hist Gradient Boosting Classifier, LGBM Classifier, Decision Tree Classifier, Ada Boost Classifier, SGD Classifier, and K neighbors Classifier. To conduct this study, we classified cases of malaria using data obtained from chest X-ray images. A dataset was created using 1079 patient records, and 23 attributes were used. These attributes were gathered from the Kaggle repository. Out of these 23 attributes, 80% of the data were used to train the model, and the remaining 20% were used to assess the validation accuracy. It has been shown that the Gaussian NB models were the most accurate, with a 97.66% accuracy rate.

TABLE OF CONTENTS

CONTENTS **PAGE**

CHAPTER

CHAPTER 3: Methodology **8-15**

CHAPTER 4: Result and discussion **16-22**

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

According to the World Health Organization (WHO), there were approximately 228 million malaria cases globally in 2018, with the disease ultimately responsible for the deaths of 405,000 people [1]. As a result, malaria is a serious worldwide health problem and a leading cause of death [2, 3]. It's a parasitic illness brought on by protozoan parasites that only have one cell. Only six Plasmodium species are known to infect humans routinely, but there are more than 120 species that infect mammals, birds, and reptiles [4]. These six species are P. vivax, P. falciparum, P. malariae, P. ovale curtisi, P. ovale wallkeri, and P. knowlesi. P. falciparum is the most common and dangerous form of the malaria parasite, responsible for an estimated 99.7% of cases in Africa in 2018 [1]. It can even be lethal in some situations. Malaria vectors are female Anopheles mosquitoes that have been bitten by a Plasmodium-infected mosquito. Over 100 nations and territories in the tropics and subtropics pose a threat of transmission [2,6]. Malaria is spread to places where it is not naturally found, like Europe, because over 125 million tourists visit these regions yearly [2]. A total of 120 to 180 new cases are reported annually in Spain [6]. The complex lifecycle of the Plasmodium parasite requires not one but two hosts, one of which is an insect vector (mosquito) and the other a vertebrate host (human) [7]. The exoerythrocytic cycle begins after sporozoites (Plasmodium's infective stage) are introduced into the human circulatory system. The parasite replicates within the liver cells and then releases merozoites. When liver cells die, the erythrocytic cycle begins. Erythrocytes, often known as "red blood cells," are the most numerous type of cell in the blood, making up between 40 and 45 percent of the fluid. Platelets and white blood cells are the other two types of cells that float around in blood (WBCs). Ring-shaped trophozoites are the first stage of parasite development inside erythrocytes. They multiply within the RBC, transforming into schizonts until there are so many that the cell bursts. The erythrocytic cycle is repeated when they are discharged and infect new erythrocytes [6].

Accordingly, this is a laborious process because it takes pathologists 30–60 minutes to examine each patient's sample [5] carefully. Parasite confirmation in early malaria cases ensures species-specific antimalarial treatment, lowers the mortality rate [7], and suggests other illnesses in negative patients [4]. Furthermore, it requires a high level of expertise, is subjective, and is prone to error [8], and healthcare personnel may have trouble diagnosing malaria in areas where it is not endemic [3, 5].

To overcome this situation in this study we build a machine learning based identification system to detect malaria fever.In order to forecast Malaria fever, we used a number of machines and various forms of deep learning. We utilized a dataset consisting of 1079 instances of malaria that were officially recorded. In the course of our research, we have shown a variety of findings, the most prominent of which are testing, accuracy, area under the curve (AUC), sensitivity, specificity, false-positive rate, and false-negative rate. We used hyperparameter tweaking in order to get optimum performance on our data in a reasonable amount of time. This was accomplished by finding the optimal combination of hyperparameter value combinations. We also used ANN with our data, and it resulted in an accuracy rate of 97 percent for us. And with everything taken into account, we were eventually able to attain a 98 percent accuracy rate with our forecast. In addition, the Gaussian naïve Bayes technique was used in the construction of our proposed models. A discussion has also taken place over the ways in which our model is an improvement over the Mariki model (Mariki 1).

The study's most important originality is in the following areas:

- It provided a method for recognizing malaria infection based on symptoms in a patient.
- The data were classified using several different machine learning and data processing methods.
- The work contributed to society by presenting a very accurate detection method based on a massive dataset.
- Models were categorized in a way that guaranteed the fastest possible compilation with the maximum possible accuracy.
- The most efficient model was selected after a performance analysis of the models already in use.

1.1 Objectives

Our main goal is to develop a complete and usable system for predicting malarial fever prior to a clinical trial. Many people suffer from this illness, but lack of money and awareness they can't diagnosed. We want to help them by predicting malarial fever prior to a clinical trial.

1.2 Motivation

Malarial fever is one of the many illnesses that affect millions of people. According to the World Health Organization's (WHO) most recent World Malaria Report, there will be 241 million malaria cases and 627 000 malaria deaths globally in 2020. From those who death by malaria, most of them don't know that they have malaria for the lacking of diagnosed. We want to give them a better way to identify whether they have malaria or not. Because if it is not identified and take treatment in early stages, it can have fatal consequences.

1.3 Research Question

As like as research question for this research are:

1. How can we develop and test a number of systems?
2. How to maintain vast dataset for predicting malarial fever prior to clinical trial.
3. How can the dataset be fully sorted by classifying?

1.4 Challenges

There are several issues to consider during this research.

1. Identifying and testing a method for putting multiple systems into action.
2. Creating a large dataset.
3. Classifying the dataset by determining its different types and classes.

1.5 Report Organization

In this full thesis, we divided into different chapter. First, in chapter 1 we discuss the introduction and also make some sub section like objective, motivation, researchquestion, and challenges. In next section review in chapter 2. Chapter 3 discuss about methodology. In chapter 4 we describe the performance evaluation, chapter 5 about the result and discussion and lastly chapter 6 contain conclusion and future work.

CHAPTER 2

Background

2.1 Preliminaries

Machine learning is a promising tool when it comes to early sickness prediction. Research has been carried out to forecast diseases such as strokes, cancer, malaria, etc. Malaria disease is a subject that has gotten very little research.

2.2 Related works

Rosado et al. [9] described an image processing and analysis method based on a supervised classification to detect P. falciparum trophozoites and white blood cells in Giemsa stained thick blood smears. Their automatic detection of trophozoites achieved a sensitivity of 80.5 %and a specificity of 93.8 % using a support vector machine (SVM) and a mix of geometric, color, and texture features, while their white blood cell detection achieved 98.2 % sensitivity and 72.1 %specificity.

Arco et al. [10] used a Gaussian filter to extract Gaussian noise from images to detect malaria. Their study used histogram equalization, thresholding, and morphological operations methods and was conducted with 96.46% average accuracy.

R. Mopuri et al.[11] proposed a climate-based forecasting system for malaria prediction by using the SARIMA model. This model anticipates malaria cases based on the present patterns of seasonal autocorrelation in the malaria case data. According to the predicted model, malaria cases are strongly impacted by environmental conditions, particularly rainfall and temperature. The SARIMA model achieved the highest accuracy and resulted as the best fit model.

O. Nkiruka et al. [12] proposed a machine learning-based algorithm to classify malaria in six countries based on climate variability. They employed the k-means clustering method to find malaria outliers. They used the XGBoost algorithm to classify the malaria. By comparing their model to certain other models and nations, they determined that the XGBoost algorithm diagnosed more accurately, with a 98% accuracy.

T. Sajana and M. R. Narasingarao [13] proposed a majority voting algorithm for diagnosing malaria with an imbalanced data set. They analyzed their proposed method by using Naïve Bayes, Decision Tree, C4.5, and KNN classifiers and estimated that the voting model can generate 95.2% accuracy with an imbalanced dataset.

By using the Nave Bayesian algorithm, T. Sajana and M. R. Narasingarao [14] addressed an early classified model for malaria classification purposes. To get better classification results with imbalanced datasets, they used the Weka environment and R language. Their model resulted in 88.5% accuracy in the Weka environment and 87.5% accuracy with the R language..

B. Akter et al. [15] presented a machine learning model to predict brain stroke disease. In this research, they used a stroke prediction dataset from Kaggle, and it contains a total of 5110 instances. Three distinct machine learning algorithms were applied here, and Random Forest gained the maximum accuracy of 95.30%.

S. Yadav et al. [16] proposed a machine learning-based malaria prediction using clinical findings. They used SVM, Random Forest, and ANN algorithms to build their model. They obtained 92% precision, 85% recall, and an 89% F1 score for the Rapid Diagnostic Test.

A. OLUGBOJA and Z. Wang [17] developed an automated, fast, and precise system using stained blood smear images. They employed a watershed segmentation technique for plasmodium infected and non-infected samples for relevant feature extraction. Six machine learning models are utilized to evaluate the accuracy of their models and tested 99.8% of the highest TPR with Fine Gaussian SVM.

M. Mariki et al. [18] proposed a machine learning approach to diagnosing malaria depending on patient symptoms and demographic features. To train and test their model validation, they used K-fold cross-validation. They used a combined dataset by merging two datasets. Their model resulted in 95% accuracy on Kilimanjaro.

O. Iradukunda et al. [19] worked with malaria disease prediction based on 27,560 images data. SVM, KNN, CART, RF, CNN, VGG16, RESNET, and DENSENET models were applied here. The DenseNet model performed better among all, and it achieved 99% accuracy.

2.3 Comparative Study of Different Relative Research Works

Two factors of predicting malarial fever and overall accuracy can be used to describe the accuracy of a system of predicting malaria fever. Table 1 provides asummary of the methodology and issues of some earlier research projects.

Table 1: Comparative Study of Different Works

Author and year	Algorithm	Accuracy
Nkiruka, O.,Prasad,R. and Clement, O. (2021) [12]	XGBoost	86.0%
Sajana and M. R. Narasingarao (2019) [13]	Naïve Bayes	95.2%
Sajana and M. R. Narasingarao (2018) [14]	Naïve Bayes	88.5%
S. Yadav et al. (2021)[16]	Random Forest	92.0%
Mariki,M., Mkoba, E. and Mduma, N. (2022) [18]	Random Forest	95.0%

CHAPTER 3

Methodology

In order to construct the desired model, the known machine life cycle was used. Significant steps have been taken, including the feature sorting, collection of data, preprocessing, feature engineering, its division into train sets and test sets, the construction of the model, its cross-validation, hyperparameter tuning, and ultimately, the alpha test of the models.

3.1 Data Collection:

We used Google Forms to collect our desired data. Social media and Daffodil University students have been assisted in conducting our survey. Also, Data World, Google Data site, where other fever data has been found. And it has been included in the dataset as "other fever".

3.2 Data Description

We were able to collect a total of 1079 records with 23 attributes. The classes of each feature are answered with yes or no. Some missing values are created, as some questions are added later in Google form. The attributes details are presented in Table 2.

Table 2: Details of The Attributes Used in the Dataset

Feature Name	**Feature Type**	**Classes**
Age	Numerical	
Sex	Nominal	Yes=1, No=0
Headache	Nominal	Yes=1, No=0
Retro-Ocular Pain	Nominal	Yes=1, No=0
Muscle Or Muscle Joint Pain	Nominal	Yes=1, No=0
Nausea	Nominal	Yes=1, No=0
Rash	Nominal	Yes=1, No=0
Vomiting	Nominal	Yes=1, No=0
Traveler	Nominal	Yes=1, No=0

Feature Name	**Feature Type**	**Classes**
Fast Heart Rate	Nominal	Yes=1, No=0
Bloody Cough	Nominal	Yes=1, No=0
Less Urination	Nominal	Yes=1, No=0
Swollen Eyelid	Nominal	Yes=1, No=0
Sweating	Nominal	Yes=1, No=0
Nose Bleeding	Nominal	Yes=1, No=0
Shortness Of Breath - Asphyxia	Nominal	Yes=1, No=0
Sensory Change	Nominal	Yes=1, No=0
Fever With Breaks	Nominal	Yes=1, No=0
Constipation	Nominal	Yes=1, No=0
Insomnia	Nominal	Yes=1, No=0
Loss Of Appetite	Nominal	Yes=1, No=0
Muscle Stiffness	Nominal	Yes=1, No=0
Fever Type	Nominal	Malaria=1, Other Fever=0

We have collected data on patients suffering from high fever. The level of fever is variable in the affected period so we did not collect data separately for the level of fever. All of the patients at this facility had fevers with temperatures ranging from 101 to 105 degrees. During our time spent collecting data, we came across a variety of additional serious diseases in addition to Dengue fever. These infections included typhoid, chikungunya, COVID-19, and normal fever.

Which are included as other fevers in our target attribute Fever Type. Finally, the target attribute has two classes Malaria and Other Fever. So our aimed problem is a binary classification problem.

3.3 Implementation Steps of the Work

Figure 1 represents the step by step procedure of the study. The step of implementation are describe in detail in the following subsections.

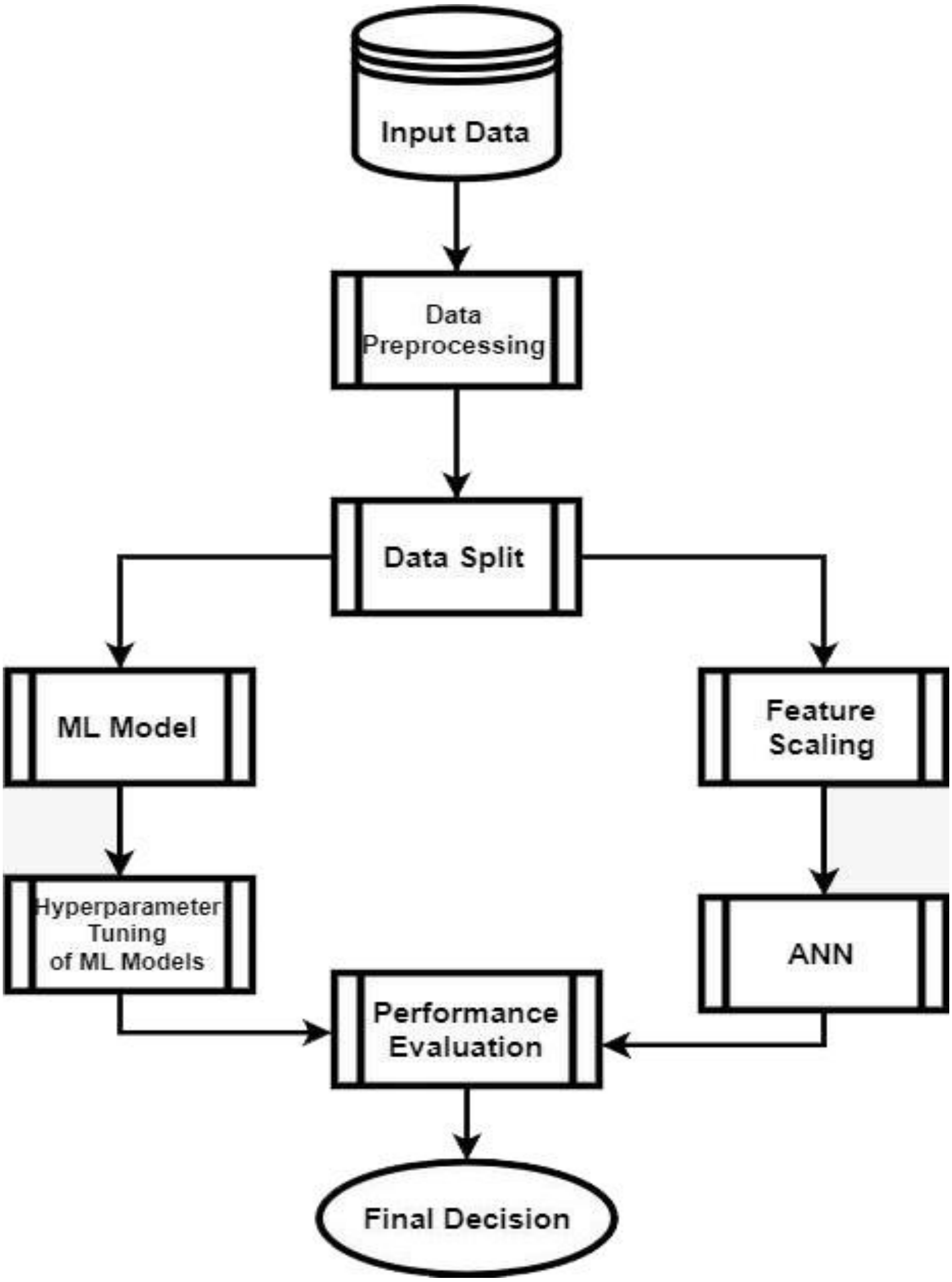

Figure 1: Step by Step Process of the Study

3.3.1 Data Preprocessing

Two days after we first started gathering data, we came across a number of new characteristics that we feel would be important to our projection. In addition, there have been reports of individuals expressing fear over the disclosure of their age. Consequently, we are missing certain data points from our collection of data. Featured that were devoid of any values, age, Nose Bleeding, Shortness Of Breath-Asphyxia, Sensory Change, Vomiting, Traveler, Swollen Eyelid, and Muscle Stiffness. These attributes have 0.8% to 1.6% missing values. Which is very low compared to dataset size and these missing values are handleable. We filled missing ages using the mean of the total age. And mode (most frequent value)is used to fill other attributes.

To build up this model we sequentially prioritized most effective featured. Datasets have nominal qualities, and the nature of those attributes is not represented in a manner that is readable by machines. Therefore, we need to transform the data into a format that can be read by machines. Utilizing the label encoding from the scikit-learn package, the value of yes has been transformed to 1, and the value of no has been changed to 0.

3.3.2 Splitting data

The dataset is split up into two sections: one for training and rest data used to check the validation of this model. To train the models, 80% of the data is utilized, whereas only 20% data is used for actual testing. Stratify guarantees that the data for all classes are spread uniformly throughout the testing and training sets.

3.3.3 Machine learning model building

Several different machine learning models [20] have been developed in order to evaluate the accuracy of predictions. Examples include the Gaussian NB, Logistic Regression, XGBoost, Bagging Classifier, Random Forest, Extra trees, Gradient Boosting Classifier, Hist Gradient Boosting, LGBM, Decision Tree, Ada Boost, SGD, and K neighbors Classifier. After the model had been developed, it was discovered that the Gaussian NB has a high level of accuracy.

3.3.3.1 Gaussian NB

The Naive Bayes classifiers are a series of straightforward "probabilistic classifiers" derived from the application of Bayes' theorem with the strong assumption that the characteristics are independent of one another. For overcome the limitations because of the redundant data and getting more pre-sized output Gaussian NB classifier utilized here. Equation 1,2 measure the performance of this classifier for suggested model. [21]

$$\mathrm{P(A \mid B)} = \frac{\mathrm{P(A \cap B)}}{\mathrm{P(B)}} \quad \text{..........} \quad 1$$

$$\text{Or } \mathrm{P(A \mid B)} = \frac{\mathrm{P(A) \cdot P(B|A)}}{\mathrm{P(B)}} \quad \text{..........} \quad 2$$

3.3.3.2 Logistic Regression

Logistic regression is a way to use statistics to predict a yes or no answer based on what has been seen in a data set before. It is a statistical technique that is used to characterize data and explain the connection between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level explanatory variables. [22]

3.3.3.3 XGBoost

Underneath the Gradient Boosting framework, XGBoost is an open-source software package that provides optimally distributed gradient boosting ml algorithms. Extreme Gradient Boosting (XGBoost) is a distributed gradient-boosted decision tree (GBDT) machine learning toolkit that is scalable. It is the top machine learning package for regression, classification, and ranking problems and offers parallel tree boosting. [23] In order to comprehend XGBoost, it is essential to first comprehend the machine learning ideas and techniques upon which it is based: supervised machine learning, decision trees, ensemble learning, and gradient boosting. Supervised machine learning uses algorithms to train a model to detect patterns in a labeled and feature-rich dataset and then uses the trained model to predict the labels on a new dataset's features.

3.3.4 Hyperparameter tuning

Each machine learning method in the sklearn package has a number of parameters with varying values. To acquire the model's ideal parameters, hyperparameter tuning is necessary. GridSearchCV is a method inside sklearn that analyzes all potential parameter values and finds which model has the best performance. [24] This is how we find the ideal model using GridSearchCV.

3.3.5 Feature Scaling

Feature Scaling is a method for standardizing the data's independent characteristics within a specified range. These operations are carried out in data pre-processing to deal with the wide range of possible magnitude or value variations that may be encountered. We used a standard scaler from sklearn to ranged attribute values. [25]

3.3.6 Used ANN Model

The word "Artificial Neural Network" originates from biological neural networks, which are accountable for the formation of the structure of the human brain during the course of evolution. In a manner analogous to that of the natural brain, which is made up of neurons connected to one another, artificial neural networks are made up of neurons connected to one another on a variety of different levels. Nodes are a collective noun that refers to these neuronal connections. [26] The Keras library was used during the construction of our neural network simulation. We improved the results and neurons by using Keras-Tuner to adjust the settings for the general model of an artificial neural network, the net input may be computed as follows:

$$Y=x1.w1+x2.w2+x3.w3.....+Bias \ldots\ldots\ldots\ldots 3$$

$$\text{Or } Y=m\ I\ xi.wi\ Z= \text{Activation}(Y) \ldots\ldots\ldots\ldots 4$$

The activation function typically activates the neurons. The RELU and Sigmoid models were used for the prediction in our test.

Currently ANN is playing an important role in predictive analysis in medical science. So we have reviewed the results of the ANN model in the case of our predictive model. We extracted the best neurons and layer numbers using Keras Tuner. Other configurations are activation = relu, output layers activation = sigmoid, loss = binary_crossentropy, learning_rate: 0.0001. Layers and neurons are presented in Table 3.

Table 3: Properties of the Used ANN Model

Layers	**Neurons**
Unit 0	19
Unit 1	224
Unit 2	352
Unit 3	416
Unit 4	160
Unit 5	192
Unit 6	224
Unit 7	352
Unit 8	160
Unit 9	256
Unit 10	384
Unit 11	32
Unit 12	32
Unit 13	32
Unit 14	32
Unit 15	32
Unit 16	32
Unit 17	32

Layers	**Neurons**
Unit 18	32

3.3.7 Performance Evaluation

In this step we have evaluated the performance of our models using different performance metrics.

3.3.8 Final Decision

Based on the performance metrics, we have selected the best model.

3.4 Cross-Validation

Cross-validation is used as a preventive technique against overfitting in prediction models when there is a restricted quantity of data available for analysis. This is of the utmost significance. The statistical technique known as cross-validation requires creating a certain number of folds, or divisions, of the data, which is then followed by an analysis of each fold and an arithmetical averaging of the estimation of the total amount of error. In order to carry out the procedures involved in the cross-validation testing, ten duplicates of our data are produced and used. In addition, the outcomes are emphasized in the section labeled "Results".

CHAPTER 4
Results and Discussions

We anticipated developing a model via the use of machine learning algorithms that, in the absence of any clinical data, would provide a concept of malaria fever. With the assistance of a variety of distinct machine learning methods, a variety of distinct sorts of prophecy models have been developed. It is vital to establish, on the basis of their findings, which model provides the highest level of performance.

4.1 Confusion Matrix

The confusion matrix for different algorithms are represented in table 4.

Table 4: Confusion Matrix of Different Algorithm

Algorithms	**TP**	**FP**	**FN**	**TN**
Gaussian NB	384	6	14	452
Logistic Regression	384	6	16	450
XGB Classifier	384	6	17	449
Bagging Classifier	384	6	17	449
Random Forest Classifier	382	8	17	449
Extra Trees Classifier	381	9	17	449
Gradient Boosting Classifier	380	10	16	450
Hist Gradient Boosting Classifier	379	11	17	449
LGBM Classifier	379	11	17	449
Decision Tree Classifier	377	13	16	450
Ada Boost Classifier	372	18	20	446

4.2 Performance Evaluation

After the machine learning and deep learning models have been developed, we are put through a series of tests to see whether or not our predictions are applicable. Our model has been tested, and the results have been analyzed for its accuracy, F1 score, precision, recall, cross-validation, and interpretability.

4.2.1 Accuracy

The evaluation of the appropriateness of the examination is based on pictures that were not shown during the practice sessions. It has not been demonstrated that a model based on a technique is accurate. Financially speaking, it won't amount to much at all.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots .5$$

4.2.2 Precision

Precision is measured by the number of correctly anticipated positive categorizations produced by a model, regardless of whether the classifications are proper. To do this, the following formula should be used:

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots .6$$

4.2.3 F1 score

Because Precision and Recall [27] are averaged in the F1 score, it provides an all-encompassing perspective of these two variables. This is why the F1 score is so important. We will get the highest possible F1 score when both Precision and Recall are correct.

$$\text{F1Score} = \frac{2 \times (\text{Recall} \times \text{Precision})}{(\text{Recall} + \text{Precision})} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 7$$

4.2.4 Recall

The recall metric determines how well a model may be anticipated after being shown to the user. The number of solid courses is not quite up to the standard of perfection that earned fantastically. If possible, it ought to be as complete as is humanly possible. [27]

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 8$$

4.2.5 Area Under the Curve (AUC)

Area Under the Curve is what the acronym AUC [28] refers to when written out. It is used as a performance indicator over a broad range of thresholds, and there are many different kinds of these. It indicates the degree to which two things may be differentiated. In addition, it demonstrates how effectively the model can distinguish between the many data categories that may be used.

4.2.6 Specificity

In the context of an unaffectedly negative model, the true negative rate, sometimes referred to as specificity, [29] is the percentage of samples that test negative when the test is used. In other words, it is the rate at which the test is successful in excluding positive results.

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 9$$

4.2.7 False Positive Rate (FPR) and False Negative Rate (FNR)

The expectation of the false positive ratio is typically referred to as the false positive rate (FPR).

$$\text{FPR} = \frac{\text{FP}}{(\text{FP} + \text{TN})} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 10$$

The percentage of positive test results that result in a false negative is known as the false negative rate (FNR).

$$\text{FNR} = \frac{\text{FN}}{(\text{FN} + \text{TP})} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 11$$

4.2.8 Performance of the Algorithms

The table 5 shows the performance metrics of the different algorithms used in this study.

Table 5: Performance Metrics of Different Algorithms

Algorithms	**Accuracy**	**F1 Score**	**Precision**	**Recall**	**AUC**	**Cross Validation**
Gaussian NB	97.66%	97.53%	97.46%	97.62%	98.00%	98.03%
Logistic Regression	97.42%	97.06%	969.9%	97.17%	97.70%	97.94%
XGB Classifier	97.31%	97.30%	97.22%	97.41%	97.73%	98.22%
Bagging Classifier	97.31%	97.30%	97.22%	97.41%	98.16%	98.31%
Random Forest Classifier	97.07%	97.06%	96.99%	97.15%	98.26%	97.57%
Extra Trees Classifier	96.96%	96.94%	96.88%	97.02%	97.95%	97.38%
Gradient Boosting Classifier	96.96%	96.94%	96.89%	97.00%	97.59%	97.85%
Hist Gradient Boosting Classifier	96.72%	96.71%	96.66%	96.77%	97.66%	96.73%
LGBM Classifier	96.72%	96.71%	96.66%	96.77%	97.53%	96.82%
Decision Tree Classifier	96.61%	96.59%	96.56%	96.62%	96.62%	96.82%
Ada Boost Classifier	95.91%	95.88%	95.89%	95.87%	97.04%	97.75%
SGD Classifier	92.99%	92.98%	93.14%	93.45%	98.00%	96.35%
K neighbors Classifier	88.55%	88.32%	89.13%	88.00%	92.90%	96.81%

The values of our tests, as well as a comparison of the different methods, are shown in Table 5.

The efficiency of the algorithms is pretty comparable to one another. On the other hand, the Gaussian NB method seems to have the greatest performance when measured by accuracy. It has an accuracy rating of 97.66 percent overall. Additionally, 97 percent accuracy can be achieved with Logistic Regression, XGB Classifier, Bagging Classifier, and Random Forest Classifier. These algorithms also demonstrate the legitimacy of their use. Let's do some deep analysis to determine which of these models is the most effective.

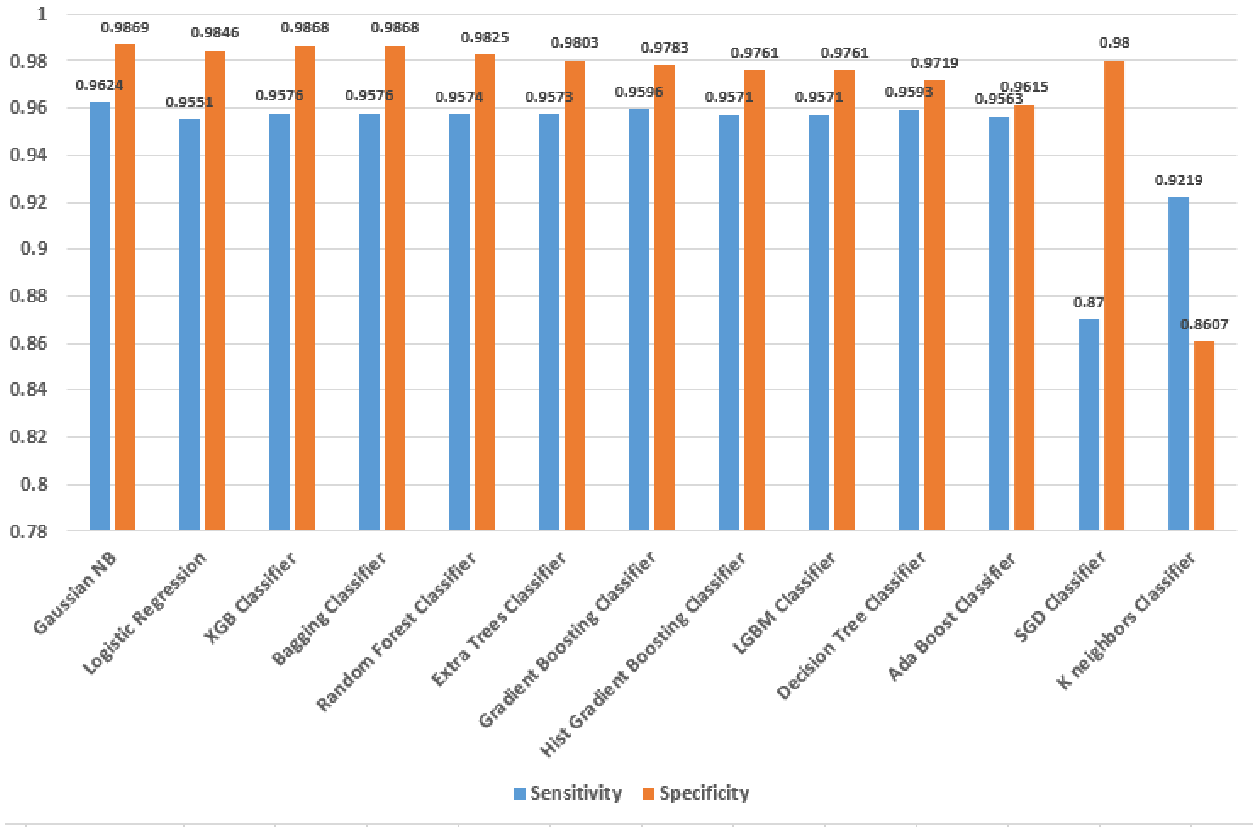

Figure 2: Visualization of Sensitivity and Specificity

Figure 2 specifies which algorithm we choose best. In Figure 2 we can see that Gaussian NB is the only model whose sensitivity can cross 0.96. The quality of specification is in a satisfactory position. Also Figure 3 shows that Gaussian nb's FNR is the lowest. All other models have values above 0.04, but Gaussian NB is below 0.4. These measurements put the Gaussiannb model ahead of other models.

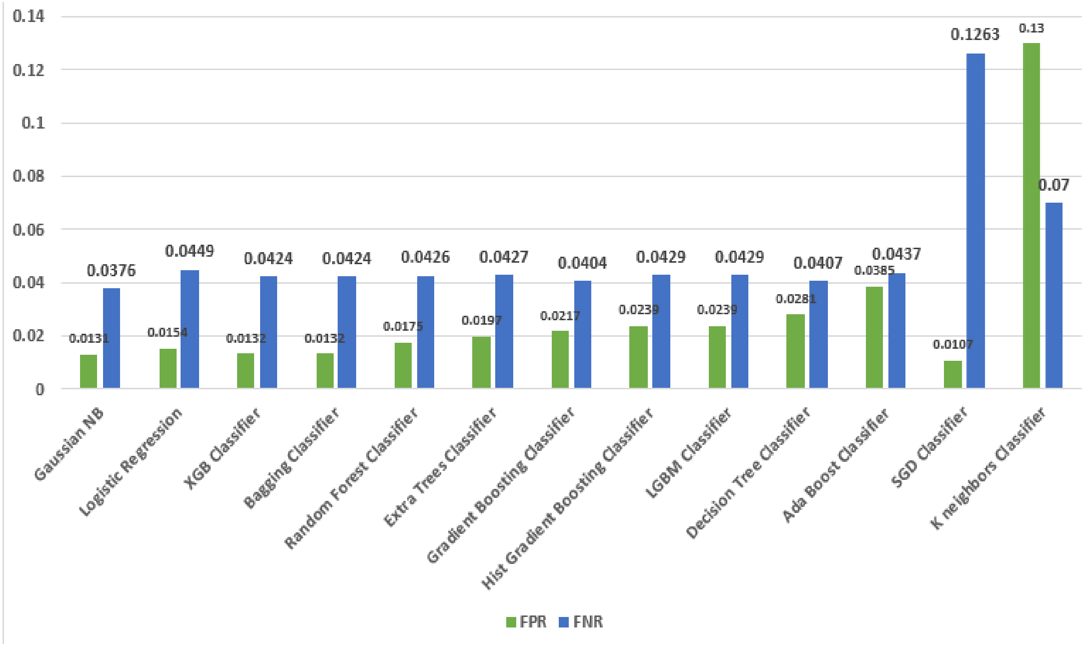

Figure 3: False Positive Rate (FPR) and False Negative Rate (FNR)

4.2.9 Visualization of ROC Curve

A receiver operating characteristic (ROC) curve is a graphical representation of the total classification levels of a categorization model. [28] This curve illustrates the relationship between two different variables: the True Positive Rate and the False Positive Rate.

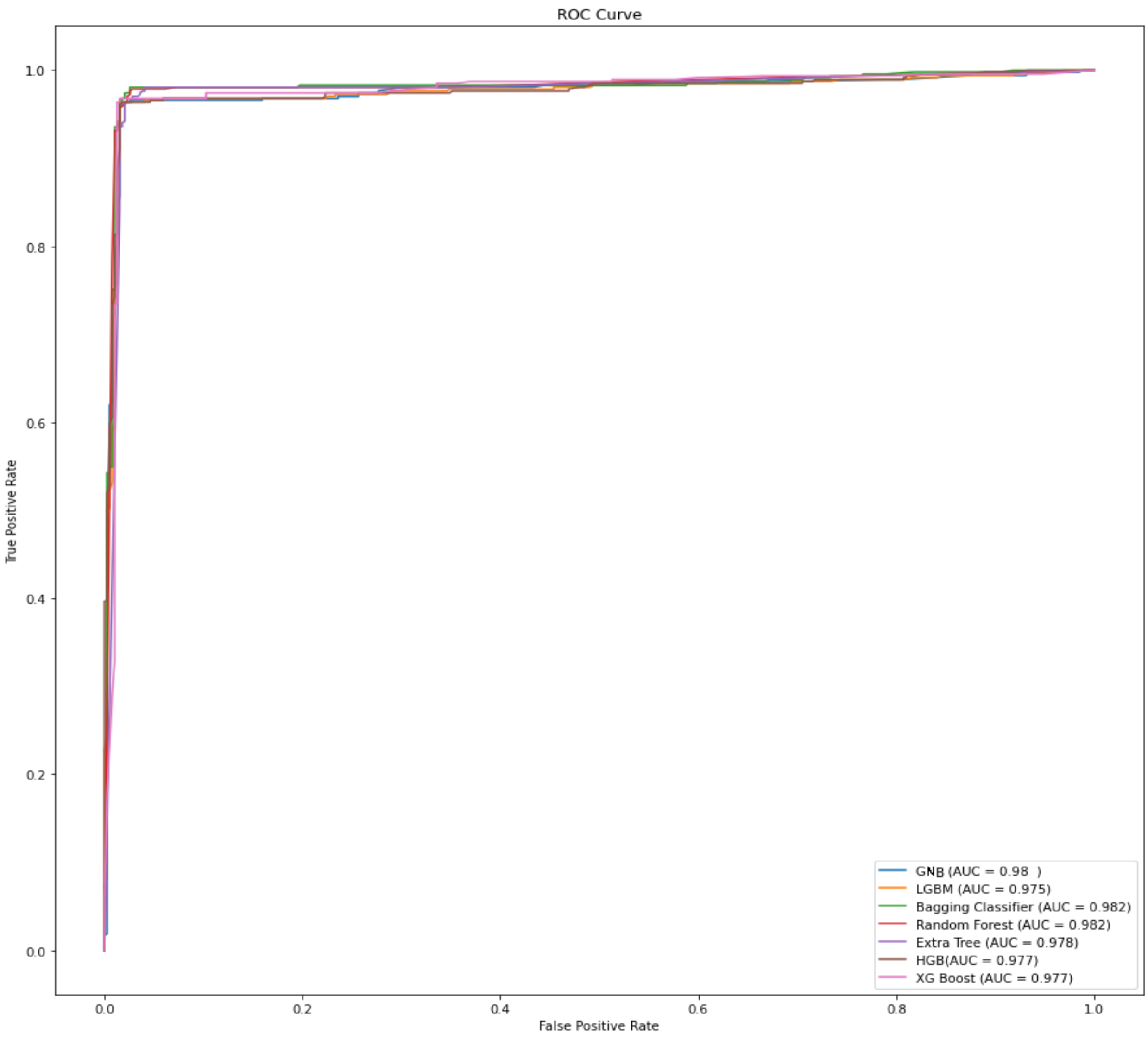

Figure 4: Model ROC curve

When we take a look at the roc curve in Figure 4, we see that the Gaussian NB, Bagging, and Random Forest Classifier models cover around 98 percent of the available data. Even in this regard, it is difficult to determine which model is superior. From this we understand, Gaussian NB is at the forefront of classifying each class individually. These measurements put the Gaussian NB model ahead of other models.

CHAPTER 5
Impact on Society And Environment

5.1 Impact on Society

With our project a person can predict weather he/she have malaria or not. Our aim to aware people by predicts malaria fever prior to clinical trial as if they take proper treatment in time. Some might argue that not knowing what lies ahead is better for people. That is something that we vehemently disagree with because if people knew they might prepare and do what was important to them.

5.2 Impact on Environment

Due to the fact that we have not yet fully developed a carbon neutral eco-system, all developing nations have a higher carbon footprint than any other nations. Our project, which uses cloud-based software, does very little to have a wide range of environmental effects. We only want to empower people with the knowledge they need to advance independent causes. Consequently, this project does not directly cause any harm.

5.3 Ethical Aspect

As previously stated, we are working on an ML-based project. There is no algorithm that can predict anything with 100% accuracy. As a result, our findings are not facts but rather possibilities. If anything is unethical, the entire system is flawed, which is not the case. The majority of the time. Similarly, our project has no negative ethical implications.

CHAPTER 6
Conclusions and Future work

6.1 Conclusions

The detection of malaria illness prior to clinical testing using machine learning and deep learning algorithms is the primary objective of this research article. We drew our conclusions about the outcomes of our experiment based on a comparison of the findings presented in a variety of studies that used a variety of methodologies. Some have attempted to predict Malaria using data from clinical tests. Their model had the greatest results, with an accuracy ranging from 92 percent to 95 percent. We obtained a success rate of 98 percent by employing the Gaussian NB Classifier rather than the non-clinical data model. When it comes to the prediction of malaria, we have used a radically different approach. In the part on the outcomes, we went through how our road is unique in comparison to others, and how our suggested model is doing very well in spite of this. It is hoped that by doing this, medical professionals will be able to assess Malaria by using our model and will be able to prescribe appropriate clinical testing, which would result in a reduction in the cost to patients.

6.2 Future work

The performance of our model may be increased if we application different feature selection technique. In future we will applied our model for identification different types disease with different merged dataset.

References

[1] World Health Organization. World Malaria Report 2019; World Health Organization: Geneva, Switzerland, 2019; Licence: CC BY-NC-SA 3.0 IGO.

[2] World Health Organization. International Travel and Health: Situation as on 1 January 2010; World Health Organization: Geneva, Switzerland, 2010.

[3] Askling, H.H.; Bruneel, F.; Burchard, G.; Castelli, F.; Chiodini, P.L.; Grobusch, M.P.; Lopez-Vélez, R.; Paul, M.; Petersen, E.; Popescu, C.; et al. Management of imported malaria in Europe. Malar. J. 2012, 11, 328. [CrossRef] [PubMed]

[4] EA, A., 2018. Pyae Phyo 2, Woodrow CJ. Malaria. *Lancet*, *391*(10130), pp.1608-1621.

[5] Tangpukdee, N., Duangdee, C., Wilairatana, P. and Krudsood, S., 2009. Malaria diagnosis: a brief review. *The Korean journal of parasitology*, *47*(2), p.93.

[6] Merino, A., 2019. *Manual de citología de sangre periférica y líquidos biológicos*. Panamericana.

[7] Das, D.K., Ghosh, M., Pal, M., Maiti, A.K. and Chakraborty, C., 2013. Machine learning approach for automated screening of malaria parasite using light microscopic images. *Micron*, *45*, pp.97-106.

[8] Loddo, A., Di Ruberto, C. and Kocher, M., 2018. Recent advances of malaria parasites detection systems based on mathematical morphology. *Sensors*, *18*(2), p.513.

[9] Rosado L, da Costa JMC, Elias D, Cardoso JS. Automated detection of malaria parasites on thick blood smears via mobile devices. Procedia Computer Science. 2016;90:138–144.

[10] Arco, J.E., Górriz, J.M., Ramírez, J., Álvarez, I. and Puntonet, C.G., 2015. Digital image analysis for automatic enumeration of malaria parasites using morphological operations. *Expert Systems with Applications*, *42*(6), pp.3041-3047.

[11] Mopuri, R., Kakarla, S.G., Mutheneni, S.R., Kadiri, M.R. and Kumaraswamy, S., 2020. Climate based malaria forecasting system for Andhra Pradesh, India. *Journal of Parasitic Diseases*, *44*(3), pp.497-510.

[12] Nkiruka, O., Prasad, R. and Clement, O., 2021. Prediction of malaria incidence using climate variability and machine learning. *Informatics in Medicine Unlocked*, *22*, p.100508.

[13] Sajana, T. and Narasingarao, M.R., 2018, May. Majority voting algorithm for diagnosing of imbalanced malaria disease. In *International Conference on ISMAC in Computational Vision and Bio-Engineering* (pp. 31-40). Springer, Cham.

[14] Sajana, T. and Narasingarao, M.R., 2018. Classification of imbalanced malaria disease using naïve bayesian algorithm. *International Journal of Engineering & Technology*, *7*(2.7), pp.786-790.

[15] B. Akter, A. Rajbongshi, S. Sazzad, R. Shakil, J. Biswas and U. Sara, "A Machine Learning Approach to Detect the Brain Stroke Disease," 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), 2022, pp. 897-901, doi: 10.1109/ICSSIT53264.2022.9716345.

[16] Yadav, S.S., Kadam, V.J., Jadhav, S.M., Jagtap, S. and Pathak, P.R., 2021, March. Machine learning based malaria prediction using clinical findings. In *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)* (pp. 216-222). IEEE.

[17] Olugboja, A. and Wang, Z., 2017, July. Malaria parasite detection using different machine learning classifier. In *2017 International Conference on Machine Learning and Cybernetics (ICMLC)* (Vol. 1, pp. 246-250). IEEE.

[18] Mariki, M., Mkoba, E. and Mduma, N., 2022. Combining Clinical Symptoms and Patient Features for Malaria Diagnosis: Machine Learning Approach. *Applied Artificial Intelligence*, pp.1-25.

[19] O. Iradukunda, H. Che, J. Uwineza, J. Y. Bayingana, M. S. Bin-Imam and I. Niyonzima, "Malaria Disease Prediction Based on Machine Learning," 2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP), 2019, pp. 1-7, doi: 10.1109/ICSIDP47821.2019.9173011.

[20] C. Janiesch, P. Zschech, and K. Heinrich, 'Machine learning and deep learning', *Electronic Markets*, vol. 31, no. 3, pp. 685–695, 2021.

[21] F. Itoo, S. Singh, and Others, 'Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection', International Journal of Information Technology, vol. 13, no. 4, pp. 1503–1511, 2021.

[22] Phadikar, S., Sil, J. and Das, A.K., 2012. Classification of rice leaf diseases based on morphological changes. International Journal of Information and Electronics Engineering, 2(3), pp.460-463.

[23] S. Ramraj, N. Uzir, R. Sunil, and S. Banerjee, 'Experimenting XGBoost algorithm for prediction and classification of different datasets', International Journal of Control Theory and Applications, vol. 9, no. 40, 2016.

[24] L. Yang and A. Shami, 'On hyperparameter optimization of machine learning algorithms: Theory and practice', Neurocomputing, vol. 415, pp. 295–316, 2020.

[25] G. Shobana and K. Umamaheswari, ‘Prediction of Liver Disease using Gradient Boost Machine Learning Techniques with Feature Scaling’, in 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1223–1229.

[26] I.-H. Yang, M.-S. Yeo, and K.-W. Kim, ‘Application of artificial neural network to predict the optimal start time for heating system in building’, Energy conversion and Management, vol. 44, no. 17, pp. 2791–2809, 2003.

[27] H. Huang and J. S. Bader, ‘Precision and recall estimates for two-hybrid screens’, Bioinformatics, vol. 25, no. 3, pp. 372–378, 2009.

[28] S. Narkhede, ‘Understanding auc-roc curve’, Towards Data Science, vol. 26, no. 1, pp. 220–227, 2018.

[29] T.-W. Loong, ‘Understanding sensitivity and specificity with the right side of the brain’, Bmj, vol. 327, no. 7417, pp. 716–719, 2003.

report

ORIGINALITY REPORT

23%	16%	13%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	3%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
3	www.mdpi.com Internet Source	2%
4	Submitted to University of Computer Studies Student Paper	2%
5	Submitted to Coventry University Student Paper	1%
6	lhncbc.nlm.nih.gov Internet Source	1%
7	www.hindawi.com Internet Source	1%
8	Submitted to University of Southampton Student Paper	1%
9	Odu Nkiruka, Rajesh Prasad, Onime Clement. "Prediction of Malaria Incidence using Climate	1%