Malaria Disease Prediction using Machine Learning

Tania Margret Sebastian
PG Scholar
Department of Computer Application
Amal Jyothi College of Engineering Kanjirappally, India
taniamargretsebastian2022b@mca.ajce.in

Ms. Gloriya Mathew
Asst.Professor
Department of Computer Application
Amal Jyothi College of Engineering Kanjirappally, India
gloriyamathew@amaljyothi.ac.in

***Abstract* — The application of Machine learning will keep on resulting, particularly in the field of computerized diagnostics and estimating, because malaria is a significant general medical issue around the world, and infectious prevention requires quick and exact determination. Precisely It's still difficult to tell the difference between malaria and other diseases. Here the analysis of bloodstream indices can be used to help identify potential malaria cases for further investigation. As a novel paradigm for precision medicine, we intended to categorise machine learning (ML) algorithms capable of accurately predicting nMI, UM, and severe malaria (SM) in the bloodstream variables. The results demonstrate that Random Forest is promising and that it provides the optimum blend of precision, recall, and F1-score correctness results on datasets where they beat the Rapid Diagnostic Test**

Keywords—,non-malarial infections (nMI) , uncomplicated malaria (UM), Random Forest,Rapid Diagnostic Test(RDT)

I. INTRODUCTION

Malaria is a blood illness that has an immense problem all over the world. Fever, tiredness, vomiting, and headaches are some of the symptoms. If not treated swiftly and efficiently, it might result in death. The important basis for malaria detection is microscopy, which takes substantial training, however rapid diagnostic tests (RDTs) Because of their ease of use for the identification, have supplanted microscopy as a high-quality diagnostic method for malaria. RDTs have a flaw in that they can cause target antigen gene deletions. As a result, enhanced and supplementary malaria detection methods with the ability to overcome any or all of these limits are required. The goal of this study is to create a model that can accurately determine whether or not a person has malaria. To predict disease, need to use different machine learning techniques.

II. LITERATURE REVIEW

This paper[1] proposes the comparison of six different ML classification techniques for detecting malaria disease using various clinical findings features.

The Model[2] is to evaluate the performance of ELM with other machine learning methods on the Using the same datasets, create a trained model that can accurately forecast malaria illness.

This research paper [3]established diagnostic methods in the literature which rely on and require instructions from trained operators. The documented methodology avoids issues associated with "quick diagnostic" procedures, which are species-specific and have a high per-test cost when compared to other diagnostic methods.

Proposed a study[4] to assess the presence of RBC parasitaemia using a visual image or In a new technique to identify parasites in this disease, a colour snapshot of microscope-based stem malaria blood was taken location.

This model[5] develops an automatic method for the detection of Parasitic and Non Parasitic cells by using image processing and the Support Vector Machine algorithm

III. METHODOLOGY

The purpose of this study is to create a machine learning model that can anticipate future events whether nMI or UM and SM using the given dataset with accuracy, precision, recall, and F1-score. The dataset is divided into two training and testing data sets. To improve learning accuracy, the model should be trained with more data.

The steps that require to be followed are:

1. Data Collection
2. Data Pre-processing
3. Model Building
4. Evaluate
5. Result

A. **Data Collection:** The user has a dataset for malaria prediction, the attributes of this dataset are used to train the model. There are 2207 samples of data in the dataset. The dataset has haematological parameters that are RBC count, Haemoglobin, lymphocytes count etc. These data are used to classify and predict the result.

B. **Data Pre-processing:** This is the process that is done in every machine learning model. This process includes data where unwanted null columns are deleted. Their data transformation is done. Then data is split into training and test data for the processing. These all are done for the mode

DOI: 10.5281/zenodo.6791027
ISBN: *978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam*

C. **Machine Learning:** Machine learning(ML) algorithms were evaluated to classify and predict data. Haemoglobin and haematocrit stages had been now no longer blanketed withinside the modelling due to the fact they're used to help the prognosis of malaria.

D. **Random Forest:** Random Forest is a machine learning algorithm that can be used for many functions including regression and classification. The random forest model is developed with a large number of small decision trees, they are estimators. The random forest model joins the predictions of the estimators to generate a more accurate prediction.

IV. BUILD MODEL

The model building is the main process in the malaria disease prediction. The steps involved in the model building are:

1. Import the packages

```
import pandas as pd          #read,explore & clean da
import numpy as np           # data manipulation
import matplotlib.pyplot as plt   # data visualizat
import seaborn as sns             # data visualizat
```

2. Importing the dataset

```
[ ]  from google.colab import files
     uploaded = files.upload()
```

Choose Files No file chosen Upload wid

```
Saving malaria_clinical_data.csv to malaria
```

```
[ ]  df = pd.read_csv('malaria_clinical_data.csv
     df.head(7)
```

3. Find the correlation of data(fig1,fig2)

```
[ ] subset.corr()
```

	wbc_count	rbc_count	hb_level	hematocrit	mean_cell_volume	mean_corp_hb	mean_cell_hb_conc	platelet_count	platelet_distr_width	mean_platelet_vl	neutrophi
wbc_count	1.000000	-0.217584	-0.283881	-0.261736	-0.159300	-0.064770	0.066981	0.122330	0.022260	-0.100716	
rbc_count	-0.217584	1.000000	0.896557	0.822294	-0.017848	-0.348335	-0.497815	0.403424	0.001756	0.284809	
hb_level	-0.283881	0.896557	1.000000	0.969844	0.286931	0.023659	-0.302304	0.353828	0.066770	0.248323	
hematocrit	-0.261736	0.822294	0.969844	1.000000	0.279029	-0.101481	-0.479766	0.387830	0.091718	0.285169	
mean_cell_volume	-0.159300	-0.017848	0.286931	0.279029	1.000000	0.722231	-0.077781	-0.002473	0.269485	0.038404	
mean_corp_hb	-0.064770	-0.348335	0.023659	-0.101481	0.722231	1.000000	0.576017	-0.202177	0.094658	-0.189324	
mean_cell_hb_conc	0.066981	-0.497815	-0.302304	-0.479766	-0.077781	0.576017	1.000000	-0.310377	-0.167834	-0.329254	
platelet_count	0.122330	0.403424	0.353828	0.387830	-0.002473	-0.202177	-0.310377	1.000000	0.070121	-0.036159	
platelet_distr_width	0.022260	0.001756	0.066770	0.091718	0.269485	0.094658	-0.167834	0.070121	1.000000	0.233941	
mean_platelet_vl	-0.100716	0.284809	0.248323	0.285169	0.038404	-0.189324	-0.329254	-0.036159	0.233941	1.000000	

Fig(1): using data

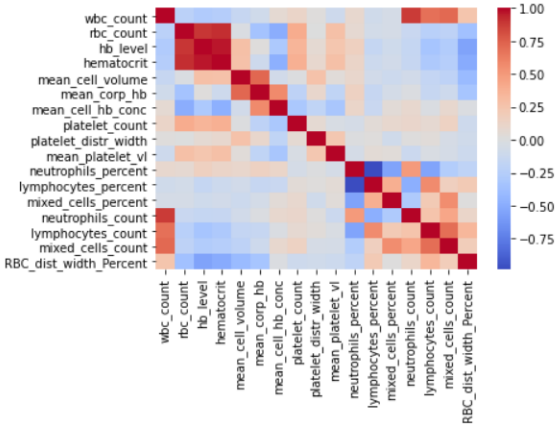

Fig2:Correlation using Heatmap

4. Spiliting the data into 80:20 ratio test and training data

```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(X,y_encoded,test_size=0.2)
```

5. Standardization of data

```
from sklearn.preprocessing import MinMaxScaler
min_max_scaler=MinMaxScaler()
X_train_scaled=min_max_scaler.fit_transform(X_train)
X_test_scaled=min_max_scaler.fit_transform(X_test)
```

```
X_train_scaled[0,0]
```

```
0.24906367041198504
```

```
X_train.iloc[0,0]
```

```
13.8
```

6. Create a random forest classifier for the training phase and model prediction on test data

Training Phase

```
#create random forest classifier
```

```
from sklearn.ensemble import RandomForestClassifier
clf=RandomForestClassifier()
clf.fit(X_train_scaled,y_train)
```

```
RandomForestClassifier()
```

Testing phase

```
# model prediction on the test set
```

```
y_pred=clf.predict(X_test_scaled)
```

```
y_pred[0:3]
```

```
array([0, 0, 2])
```

```
y_test[0:3]
```

```
array([1, 0, 2])
```

```
classes
```

```
array(['Non-malaria Infection', 'Severe Malaria', 'Uncomplicated Malaria'],
      dtype=object)
```

DOI: 10.5281/zenodo.6791027
ISBN: *978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam*

- Accuracy, Precision, Recall, and F1-score are used to evaluate the model.

```
# import the metrics
from sklearn.metrics import balanced_accuracy_score,f1_score,precision_score, recall_score
from sklearn.metrics import plot_confusion_matrix
```

```
[ ] balanced_accuracy=balanced_accuracy_score(y_test,y_pred)
    balanced_accuracy=round(balanced_accuracy,2)
    print('balanced accuracy:',balanced_accuracy)

    balanced accuracy: 0.77
```

```
[ ] f1score=f1_score(y_test,y_pred,average='weighted')
    f1score=round(f1score,2)
    print('f1score:',f1score)

    f1score: 0.74
```

```
[ ] precision=precision_score(y_test,y_pred,average='weighted')
    precision=round(precision,2)
    print('precision:',precision)

    precision: 0.76
```

```
[ ] recall=recall_score(y_test,y_pred,average='weighted')
    recall=round(recall,2)
    print('recall:',recall)

    recall: 0.75
```

V. RESULT

The result find out malaria outcomes based on the given dataset. The model is evaluated using the metrics accuracy, recall, precision, F1-score, and confusion matrix to arrive at the best result. The output is given from the model in the form of a confusion matrix. It has 2 dimensions actual and predicted.

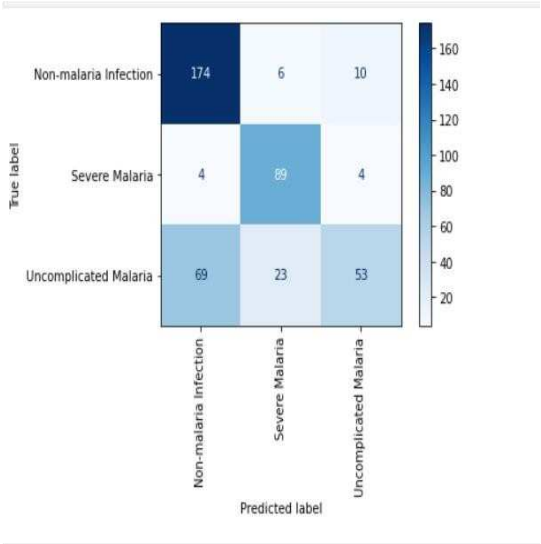

Fig3:Confusion Matrix

VI. CONCLUSION

The proposed paper is for predicting malaria disease using random forest. The haematological data for the prediction is given from the dataset. Then it was processed to get good accuracy values at the final. The model Random forest is used to predict the outcomes.

The paper future scope is the dataset with adding more data and it uses different machine learning algorithms for the prediction.

REFERENCES

[1] Samir S. Yadav, Vinod J Kadam, Shivajirao M. Jadhav, Sagar Jagtap. Machine Learning-based Malaria Prediction using Clinical Findings 2021 International Conference on Emerging Smart Computing and Informatics (ESCI).

[2] Octave Iradukunda1 , Haiying Che, Josiane Uwineza, Jean Yves Bayingana, Muhammad S Bin-Imam, Ibrahim Niyonzima Malaria Disease Prediction Based on Machine Learning. 2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP)

[3] Purwar, Y., Shah, S. L., Clarke, G., Almugairi, A., & Muehlenbachs, A. (2011). Malaria parasite detection in microscopic pictures that is automated and unsupervised. Malaria Journal, 10(1), 364.

[4] Raviraja, S., Bajpai, G., & Sharma, S. K. (2007). Analysis of detecting the malarial parasite-infected blood images using the statistical-based approach. In 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006 (pp. 502-505). Springer, Berlin, Heidelberg

[5] Usha Kumari Sanam Narejo Mashal Afzal Madeha Muzafar Memon Malaria Disease Detection Using Machine Learning.January 2021

DOI: 10.5281/zenodo.6791027
ISBN: *978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam*