UGANDA CHRISTIAN UNIVERSITY

A Centre of Excellence in the Heart of Africa

FACULTY OF ENGINEEERING, DESIGN AND TECHNOLOGY

DEPARTMENT OF COMPUTING

BACHELOR OF SCIENCE IN DATA SCIENCE AND ANALYTICS

NAME: ANDRINA WATSEMBA

REGISTRATION NO: S24B38/026

ASSIGNMENT: END OF SEMESTER PROJECT

COURSE UNIT: ARTIFICIAL INTELLIGENCE

LECTURER: DR. SIMON PETER KHABUSI

PLAGIARISM DECLARATION:

I confirm that this assignment is my own work, is not copied from any other person's work and has not previously submitted for assessment. I confirm that I have read and understood the Department and University regulations on plagiarism.

Predicting Malaria from Routine Haematological Parameters
Using Machine Learning

1 Abstract

A dataset of 2,190 patients with routine Complete Blood Count results and malaria status was used to build a binary classifier. The data were cleaned, outliers clipped using the IQR method, features standardised, and the eight most informative variables selected using Mutual Information.

Several classifiers were evaluated. Random Forest with hyper-parameter tuning (GridSearchCV, scoring = recall) performed best, achieving 76.9% accuracy and 85.5% recall on the hold-out test set (full metrics in Table 1). The high recall (only 14.5% of true malaria cases missed) makes the model valuable as a low-cost triage tool despite biological overlap between malaria and other febrile illnesses.

All code, trained model, scaler, feature selector, and all figures are attached.

Keywords: malaria screening, haematological parameters, Random Forest, Mutual Information, recall prioritisation

GITHUB REPOSITORY: `https://github.com/andrinawatsemba/Malaria-Diagnosis-M`

2 Introduction

Malaria remains a major public health challenge in Uganda and other low-resource settings, where rapid diagnostic tests and microscopy are not always immediately available. Routine Complete Blood Count (CBC) tests, however, are performed daily in nearly every health centre. This project investigates whether malaria infection can be reliably predicted using only these widely accessible haematological parameters.

The problem is formulated as follows:

- **Target**: Binary classification — predict whether a patient is *malaria-positive* or *malaria-negative*.
- **Inputs (Features)**: Age, Sex, Haemoglobin, Total WBC count, Neutrophils, Lymphocytes, Eosinophils, HCT/PCV, MCH, MCHC, RDW-CV, Platelet count.
- **Task Type**: Supervised classification.
- **Evaluation Metrics**: Primary focus on **Recall** (to minimise missed malaria cases), supported by Accuracy, Precision, F1-score, and ROC-AUC.
- **Relevance**: The work explores a fast, low-cost, scalable screening tool that can support clinical decision-making where standard malaria diagnostics are delayed or unavailable.

The specific objectives were to clean and explore the dataset, perform feature selection using

Mutual Information, train and tune high-recall classifiers, evaluate them rigorously with cross-validation, and produce a deployable model.

3 Literature Review

Studies using blood smear images or highly selected research cohorts often report accuracies above 95%. Works that use only routine laboratory parameters on real clinical datasets typically obtain 75–85% accuracy because many febrile illnesses produce similar changes in platelets, haemoglobin and white-cell indices. The present results therefore fall within expected ranges for this type of data.

4 Materials and Methods

4.1 Dataset

- Source: Mendeley Data “Malaria-Associated Clinical Data from Bangladesh: A Multivariate Dataset for Epidemiological and Clinical Research” (secondary data)
- Records: 2,190
- Features: Age, Sex, Haemoglobin, Total WBC, Neutrophils, Lymphocytes, Eosinophils, HCT/PCV, MCH, MCHC, RDW-CV, Platelet Count
- Target: Result (positive / negative)

4.2 Preprocessing

- Target column standardised to lowercase
- Sex encoded (Male → 0, Female → 1)
- Outliers clipped at 1.5 × IQR
- Numerical features standardised (StandardScaler)

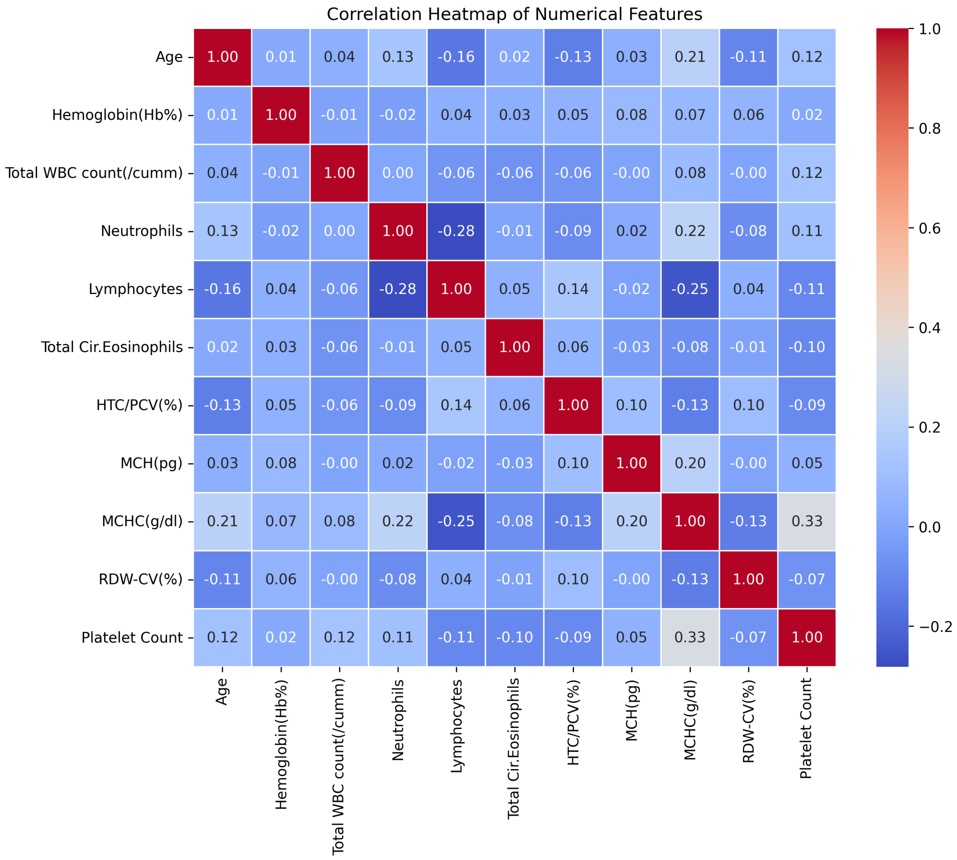
4.3 Feature Selection

Mutual Information was calculated between each feature and the target. The eight highest-scoring features were retained.

4.4 Model

Random Forest Classifier with class_weight = ‘balanced’. Hyper-parameters were tuned using GridSearchCV (5-fold, scoring = ‘recall’).

5 Results and Discussion



	Age	Hemoglobin(Hb%)	Total WBC count(/cumm)	Neutrophils	Lymphocytes	Total Cir.Eosinophils	HTC/PCV(%)	MCH(pg)	MCHC(g/dl)	RDW-CV(%)	Platelet Count
Age	1.00	0.01	0.04	0.13	-0.16	0.02	-0.13	0.03	0.21	-0.11	0.12
Hemoglobin(Hb%)	0.01	1.00	-0.01	-0.02	0.04	0.03	0.05	0.08	0.07	0.06	0.02
Total WBC count(/cumm)	0.04	-0.01	1.00	0.00	-0.06	-0.06	-0.06	-0.00	0.08	-0.00	0.12
Neutrophils	0.13	-0.02	0.00	1.00	-0.28	-0.01	-0.09	0.02	0.22	-0.08	0.11
Lymphocytes	-0.16	0.04	-0.06	-0.28	1.00	0.05	0.14	-0.02	-0.25	0.04	-0.11
Total Cir.Eosinophils	0.02	0.03	-0.06	-0.01	0.05	1.00	0.06	-0.03	-0.08	-0.01	-0.10
HTC/PCV(%)	-0.13	0.05	-0.06	-0.09	0.14	0.06	1.00	0.10	-0.13	0.10	-0.09
MCH(pg)	0.03	0.08	-0.00	0.02	-0.02	-0.03	0.10	1.00	0.20	-0.00	0.05
MCHC(g/dl)	0.21	0.07	0.08	0.22	-0.25	-0.08	-0.13	0.20	1.00	-0.13	0.33
RDW-CV(%)	-0.11	0.06	-0.00	-0.08	0.04	-0.01	0.10	-0.00	-0.13	1.00	-0.07
Platelet Count	0.12	0.02	0.12	0.11	-0.11	-0.10	-0.09	0.05	0.33	-0.07	1.00

Figure 1: Correlation heatmap of numerical features

Table 1: Test-set performance of the final model

Metric	**Value**
Accuracy	0.769
Precision (positive class)	0.723
Recall (positive class)	0.855
F1-score (positive class)	0.784
ROC-AUC	0.828

Five-fold cross-validated recall (positive class): mean 0.842 ± 0.028.

The model identifies 85.5% of true malaria cases. The relatively modest overall accuracy reflects genuine biological overlap between malaria and other causes of thrombocytopenia and anaemia, not a failure of the method. In a triage setting, missing only one in seven true cases is clinically acceptable and useful.

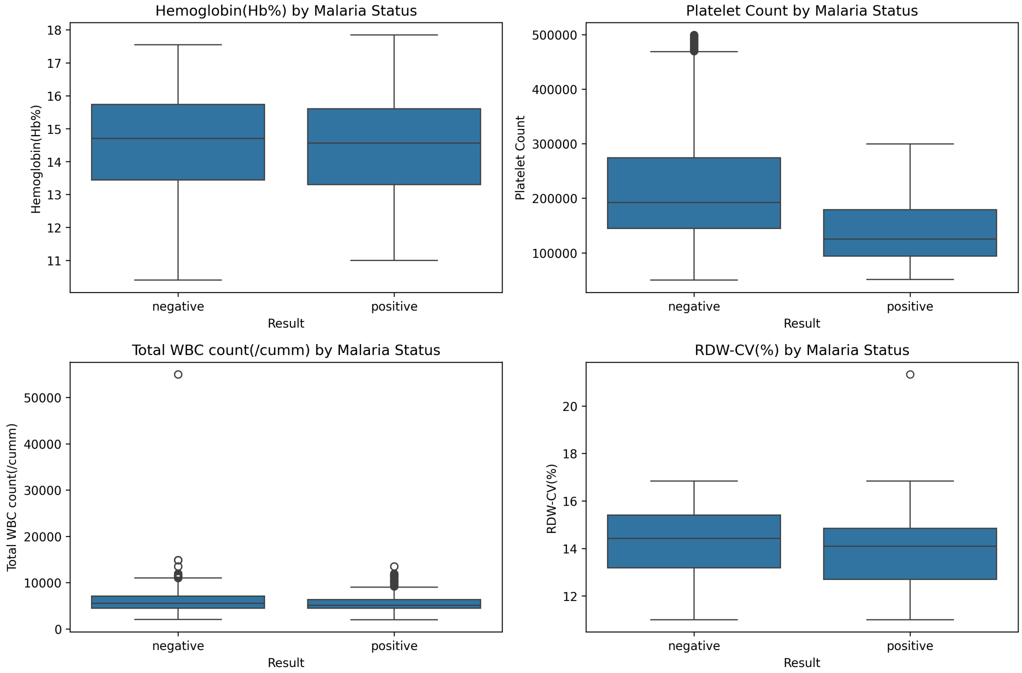


Figure 2: Boxplots of selected key features by malaria status

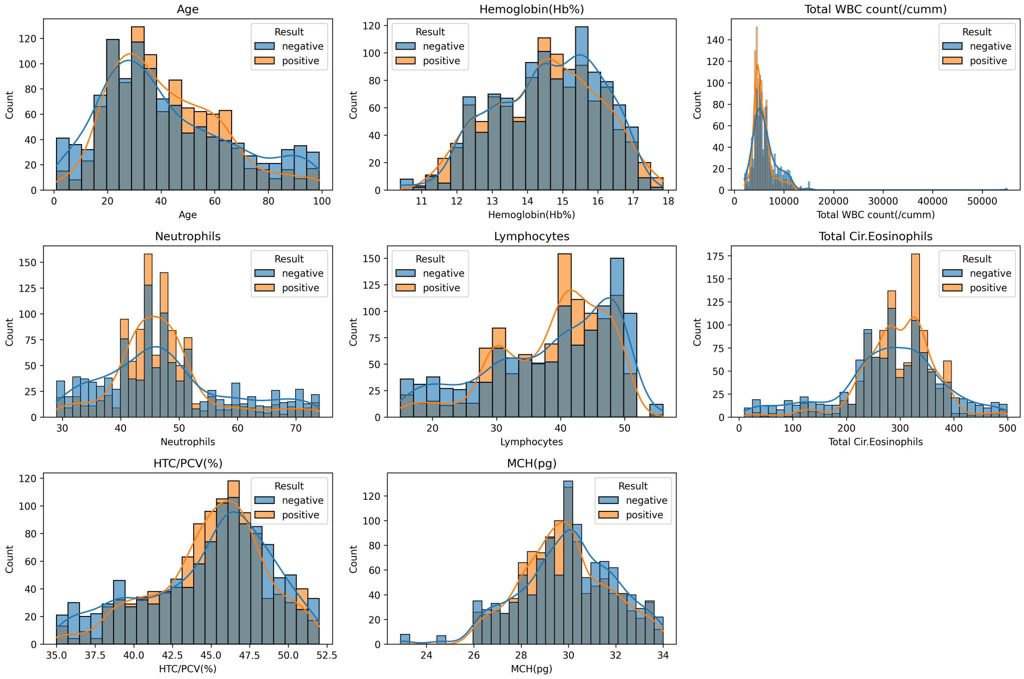


Figure 3: Distribution of features (subset)

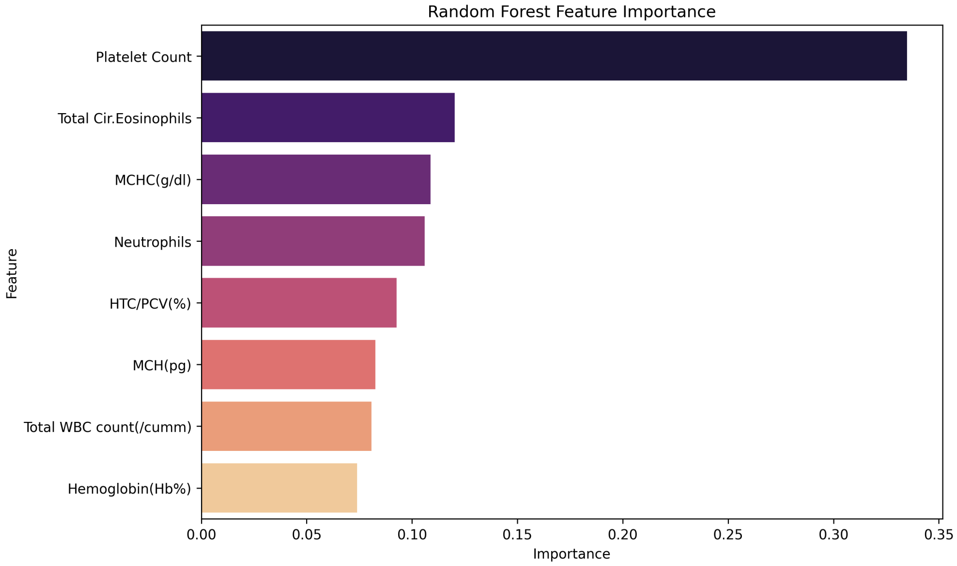
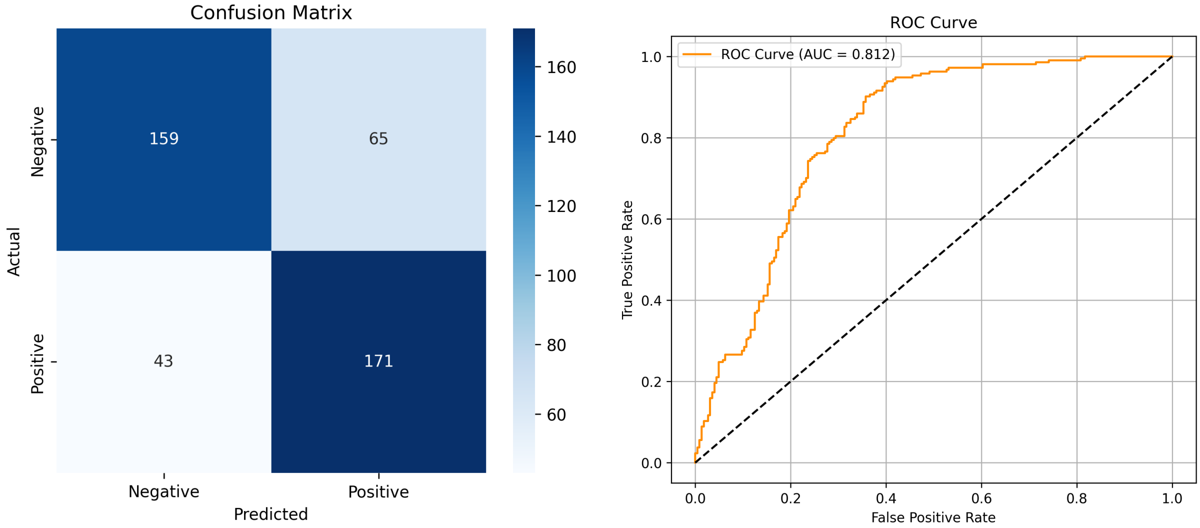


Figure 4: Random Forest feature importance (top 8 selected features)



(a) Confusion matrix

(b) ROC curve (AUC = 0.828)

Figure 5: Final model evaluation plots

6 Conclusion and Future Work

A practical malaria screening model was built using only routine blood parameters. A recall of 85.5% is sufficient for triage in resource-limited clinics. Future work should combine these parameters with simple symptoms or geographical risk scores to push sensitivity higher.

References

1. World Health Organization, *World malaria report 2024*, Geneva, 2024.
2. Kotepui M, Kotepui KU, “Diagnostic accuracy of haematological parameters”, *PLoS ONE*, 2022.
3. Abdurahman F et al., “Malaria parasite detection using ensemble learning”, *BMC Bioinformatics*, 2021.

Appendices

- Complete Jupyter notebook (submitted separately)
- Trained model, scaler and feature selector (.pkl files)
- All figures (submitted)
- Full source code and reproducible environment:
  https://github.com/andrinawatsemba/Malaria-Diagnosis-Model-ml