IJSCIA

International Journal of Scientific Advances
ISSN: 2708-7972
Volume: 7 | Issue: 6 | Nov – Dec 2026 Available Online: www.ijscia.com
DOI: 10.51542/ijscia.v7i6.15

Explainable Minimal-Feature Machine Learning for Malaria Screening Using Haematological Parameters: Secondary Analysis of a Public Bangladesh Clinical Dataset

Salma Nur Auliyya Raihanah[1*], Lilik Djuari[2]

[1]Faculty of Medicine, Airlangga University, Surabaya, Indonesia

[2]Department of Public Health, Faculty of Medicine,
Airlangga University, Surabaya, Indonesia

E-mail: salmanurauliyyar@gmail.com; lilik-d@fk.unair.ac.id
ORCID: First author 0009-0004-6011-2474; Second author 0000-0003-0872-7645

*Corresponding author details: Salma Nur Auliyya Raihanah;
salmanurauliyyar@gmail.com

ABSTRACT

Background: Routine complete blood count (CBC) parameters can support malaria screening in resource-limited areas, but the optimal minimum number of parameters required remains unclear. ***Methods:*** We performed a secondary analysis of a public dataset comprising 2,188 patients with malaria-like symptoms from Bangladesh. Using a stratified five-fold cross-validation with random forest feature ranking, we reduced 12 demographic and haematological candidate predictors down to the top four. The optimized model was evaluated on a 20% held-out test set (n = 438). ***Results:*** Platelet count, red cell distribution width, eosinophil count, and mean corpuscular haemoglobin concentration were consistently selected across all folds. On the test set, this four-predictor model achieved an area under the receiver operating characteristic curve (AUC) of 0.820 (95% CI: 0.780–0.856), sensitivity of 0.771, and specificity of 0.714. Performance was statistically equivalent to the full 12-predictor model (DeLong's test, p = 0.695), with a positive net benefit demonstrated across all decision thresholds (0.01–0.89). Shapley additive explanations identified platelet count as the dominant predictor. Four routine haematological parameters may therefore support low-cost malaria triage in resource-limited settings, pending external and prospective validation. ***Conclusion:*** A minimalist set of four routine haematological parameters can effectively support low-cost malaria triage in resource-constrained settings, pending prospective external validation.

Keywords: malaria; machine learning; feature selection; explainable artificial intelligence; haematological parameters; clinical decision support

INTRODUCTION

Malaria remains a major global health burden. The World Health Organization (WHO) estimated 282 million cases and 610,000 deaths in 2024, up from 273 million cases in 2023, with the WHO African Region accounting for most of the burden [14]. Outside Africa, transmission persists in pockets where access to diagnosis is limited. In Bangladesh, the source of the dataset analysed here, the Chittagong Hill Tracts bear more than 90% of the national malaria burden [1], and in the WHO Western Pacific Region malaria is concentrated mainly in Indonesia and Papua New Guinea [14]. As national programmes approach elimination, the remaining cases therefore cluster where diagnostic infrastructure is weakest.

Standard diagnosis relies on Giemsa-stained microscopy and histidine-rich protein 2 (HRP2)-based rapid diagnostic tests (RDTs). Both lose sensitivity at low parasitaemia, microscopy requires trained personnel, and HRP2-based RDTs are threatened by pfhrp2/3 gene deletions [13,4]. Haematological alterations, particularly a lower platelet count, are frequently observed in malaria and have been proposed as supportive rather than diagnostic indicators [12]. This is the premise of the present study: routine haematology as an adjunct screening signal that can prioritise patients for confirmatory testing, not a replacement for parasitological diagnosis.

Machine learning (ML) on routine clinical data has been explored for malaria for several years. Morang'a et al. [13] used haematological data from 2,207 Ghanaian participants and identified platelet, red blood cell and lymphocyte counts as the top classifiers distinguishing uncomplicated malaria from

non-malarial infection (AUC 0.866), explaining binary classifiers with local interpretable model-agnostic explanations (LIME). Other studies have used symptom- and demography-based classifiers in Indonesia and Tanzania [5,11], penalised logistic regression on demographic, environmental and symptom features in Nigeria [16], and explainability methods such as SHapley Additive exPlanations (SHAP), LIME and permutation importance applied to symptom-based data [2,3]. In a related disease, a publicly available Bangladeshi dengue haematology dataset has recently been analysed with a leakage-conscious tree-based framework including calibration, decision curve and SHAP analyses [24], illustrating the value of rigorous secondary analyses of routine haematological data.

To our knowledge, no malaria study has combined three elements: (i) a systematic minimal-feature reduction that asks how few routine parameters suffice, rather than ranking all available features, (ii) leakage-safe validation, in which a held-out test set is removed before feature selection and feature rankings are recomputed within each training fold, and (iii) a formal statistical comparison between the reduced and the full-feature model against a pre-specified margin. This matters in practice, because a model requiring fewer inputs is easier to deploy where a basic complete blood count (CBC) is available but malaria-specific diagnostics are not.

We therefore asked whether a minimal subset of routine haematological parameters can support malaria screening with performance equivalent to that of the full parameter panel while remaining interpretable. We pre-specified three hypotheses: (H1) a subset of four or fewer parameters performs equivalently to the full 12-parameter model, defined as an AUC difference within ±0.05, (H2) platelet count dominates model output, consistent with the known malaria–thrombocytopenia association, and (H3) the resulting model achieves a held-out AUC of at least 0.80, our minimum for a screening (not diagnostic) tool.

METHODS

1. Data Source and Study Design

We conducted a retrospective secondary analysis of a publicly available, de-identified dataset, Malaria-Associated Clinical Data from Bangladesh [19], deposited in Mendeley Data (version 4, August 2025). The dataset contains 2,190 patients with malaria-like symptoms attending healthcare centres in Bangladesh and a laboratory-confirmed binary malaria outcome. The 12 candidate predictors were age, sex, haemoglobin, total white blood cell (WBC) count, neutrophil and lymphocyte percentages, absolute eosinophil count, haematocrit/packed cell volume (HCT/PCV), mean corpuscular haemoglobin (MCH), mean corpuscular haemoglobin concentration (MCHC), red cell distribution width–coefficient of variation (RDW-CV) and platelet count (Table 1). To our knowledge, the dataset had not previously been analysed with ML. Reporting follows the TRIPOD+AI statement [16], adapted for a secondary, exploratory analysis.

TABLE 1: Candidate Predictors and Outcome in The Analysed Dataset.

Variable	Description	Unit/Coding
Age	Patient age	Years
Sex	Patient sex	Male = 1 / female = 0
Haemoglobin	Haemoglobin concentration	g/dL
Total WBC count	Total white blood cell count	Cells/μL*
Neutrophils	Neutrophils in differential count	%
Lymphocytes	Lymphocytes in differential count	%
Eosinophil count	Absolute circulating eosinophil count	Cells/μL
HCT/PCV	Haematocrit / packed cell volume	%
MCH	Mean corpuscular haemoglobin	pg
MCHC	Mean corpuscular haemoglobin concentration	g/dL
RDW-CV	Red cell distribution width, coefficient of variation	%
Platelet count	Platelet count	Cells/μL*
Result (outcome)	Laboratory-confirmed malaria test result	Positive = 1/ negative = 0

WBC: white blood cell; HCT/PCV: haematocrit/ packed cell volume; MCH: mean corpuscular haemoglobin; MCHC: mean corpuscular haemoglobin concentration; RDW-CV: red cell distribution width (coefficient of variation).

*Labelled $\times 10^3/\mu L$ in the repository file, although the recorded values are on a per-μL scale (e.g., median WBC ≈ 5,800). Column labels were corrected to cells/μL before analysis.

2. Data Quality Assessment and Preprocessing

The raw dataset was audited before analysis. It contained no missing values and no duplicate records, so the pre-specified missing-data policy (exclusion of predictors with more than 50% missing values and complete-case analysis without imputation) removed no data. Three issues were found: (i) one inconsistently capitalised outcome label, which was case-normalised, (ii) two records with an implausible WBC count of 55,000 cells/μL (reference range ≈ 4,000–11,000), identified by a pre-specified exclusion rule of WBC > 20,000 cells/μL and excluded rather than imputed, and (iii) 190 records (8.7%) in which neutrophil plus lymphocyte percentages exceeded 100%, which is physiologically implausible for a differential count.

These 190 records were retained in the primary analysis, and their influence was quantified in a pre-specified robustness check (Sect. 2.9). The analytic dataset comprised 2,188 patients (1,120 negative, 1,068 positive); because the classes were nearly balanced, no resampling was applied (Figure 1).

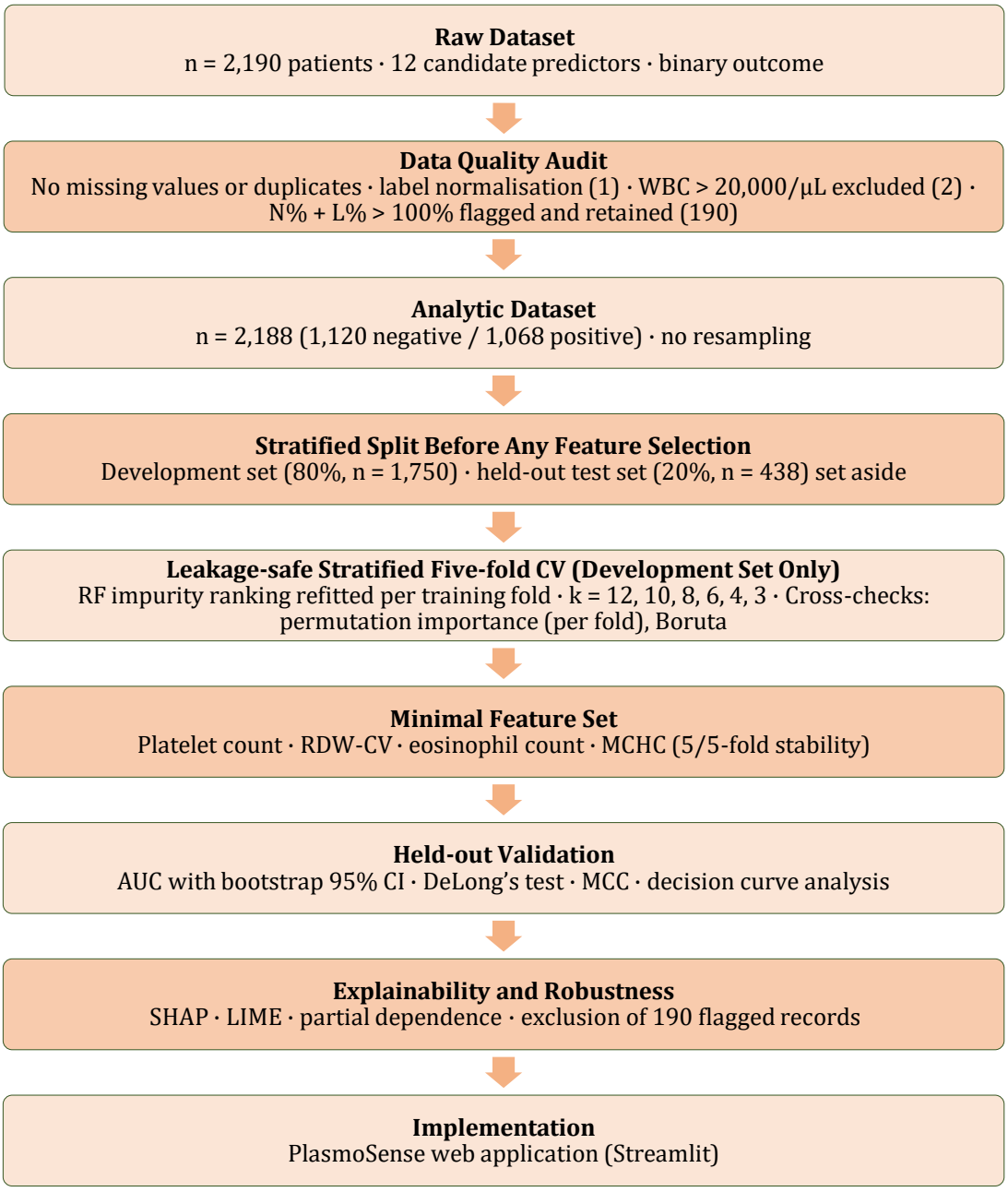


FIGURE 1: Study workflow, from the raw dataset through data-quality auditing, partitioning, leakage-safe cross-validated feature selection, held-out statistical validation and explainability analysis. CI: confidence interval; CV: cross-validation; LIME: local interpretable model-agnostic explanations; MCC: Matthews correlation coefficient; N% + L%: neutrophil plus lymphocyte percentage; RF: random forest; SHAP: SHapley Additive exPlanations.

3. Data Partitioning and Leakage-Safe Cross-Validation

Before any feature selection, the analytic dataset was split once into a development set (80%, n = 1,750) and a held-out test set (20%, n = 438), stratified by outcome with a fixed random seed (42). The held-out set was not used for feature ranking, feature-set selection, stability assessment, or model training. Ranking features on the full dataset before cross-validation lets validation information shape which features are retained, a common source of optimistic bias. All selection analyses therefore used stratified five-fold cross-validation within the development set, in which feature ranking was recomputed independently within each training fold, and the corresponding validation fold was never used for selection; metrics were averaged across folds.

4. Feature Selection and Selection Stability

Features were ranked by random forest (RF) impurity-based importance [12], with 300 trees fitted per training fold. The top k features were used to train an evaluation model within the same fold for k = 12, 10, 8, 6, 4 and 3, tracking performance against feature-set size. The candidate minimal set was defined as a priori at k = 4, and stability was assessed as the number of folds selecting an identical set.

Because impurity-based importance favours high-cardinality numerical features, we cross-checked the selected set with two methodologically distinct procedures, both restricted to the development set. Permutation importance was computed within the same fold structure by shuffling each feature in the training fold and measuring the decrease in AUC (10 repeats per feature and fold); the top four features were recorded in each fold. Boruta [9], applied once to the development set with an RF base learner (maximum depth 7, automatic number of trees, 60 iterations), compared each feature with randomly permuted "shadow" copies and confirmed a feature as relevant if it was repeatedly more important than the best shadow feature. Boruta answers a different question, namely whether a feature carries signal beyond noise rather than whether it belongs to the smallest sufficient subset, so the two procedures were treated as complementary.

5. Model Development

RF was the primary classifier because it handles non-linear relationships without feature scaling, requires little tuning, and provides the importance measures used for feature reduction. All RF models used 300 trees, default scikit-learn hyperparameters [17], and a fixed random seed (42). Hyperparameters were not tuned, so that feature-set sizes were compared under identical settings. Logistic regression (with standardised predictors) and extremely randomised trees (Extra Trees; 300 trees) were evaluated as exploratory comparators with the same partition and seed.

6. Held-out Validation and Statistical Comparison

The final four-feature model was trained on the development set and evaluated once on the held-out test set. For a formal comparison, the twelve-feature model was trained on the same development set and evaluated on the same held-out patients, yielding paired predictions. We computed percentile bootstrap 95% confidence intervals (CIs) for the AUC (1,000 resamples), DeLong's test for the AUCs of two correlated receiver operating characteristic (ROC) curves [8] and a 95% CI for the AUC difference derived from the DeLong variance, so that H1 could be assessed against the pre-specified equivalence margin (±0.05). Threshold-dependent metrics (sensitivity, specificity, positive and negative predictive values [PPV, NPV], F1-score and the Matthews correlation coefficient [MCC]) were computed at a probability threshold of 0.5. MCC was included because it summarises all four cells of the confusion matrix [19]. Calibration was assessed with the Brier score and a calibration curve based on eight quantile bins.

7. Clinical Utility Assessment

Discrimination and calibration alone do not establish clinical usefulness. We therefore performed decision curve analysis (DCA) [22], quantifying the net benefit of using the model to guide referral for confirmatory testing across threshold probabilities p_t, relative to treat-all and treat-none strategies:

$$Net\ benefit = \frac{TP}{n} - \frac{FP}{n} \times \frac{p_t}{1 - p_t} \quad (1)$$

where TP and FP are the numbers of true and false positives at threshold p_t, and n is the number of patients. A model is considered clinically useful where its net benefit exceeds that of both reference strategies. DCA was computed for both models at 45 thresholds from 0.01 to 0.89 in steps of 0.02.

8. Model Explainability

Three complementary methods based on different principles were applied to the final model, so that convergent findings would provide stronger evidence than any single method. SHAP [24], computed with TreeExplainer on the held-out patients, attributes each prediction's deviation from the average output to the individual features. Global importance was summarised as the mean absolute SHAP value, and direction as the Spearman correlation between feature values and SHAP values. LIME [22], fitted with the development set as background data and discretised continuous features, was computed for 100 randomly sampled held-out patients. Because LIME fits a local linear surrogate around each instance, we quantified SHAP–LIME agreement as the proportion of patients for whom both methods identified the same top feature, overall and separately for confidently predicted patients ($|p - 0.5| \geq 0.15$) and patients near the decision boundary ($|p - 0.5| < 0.15$). Partial dependence plots (PDPs; 25-point grid, development set) described the average marginal relationship between each feature and the predicted probability. Because PDPs show a population-average effect whereas SHAP values are instance-specific and interaction-aware, divergence between them was treated as informative rather than contradictory.

9. Robustness Analysis

To assess whether the physiologically inconsistent records (Sect. 2.2) affected performance, the five-fold cross-validated AUC of the four-feature model, with the feature set held fixed, was recomputed on the development set with and without these records.

10. Software Implementation and Reproducibility

Analyses used Python 3.12 with scikit-learn 1.9.1 [17], NumPy 2.5.3, pandas 3.0.6 and SciPy 1.18.1, shap 0.52.0 and lime 0.2.0.1 for explainability, Boruta 0.4.3 for all-relevant selection, and a fast DeLong implementation [8,21]. A fixed seed (42) was used throughout. Because RF results under a fixed seed can differ across scikit-learn releases, the full pipeline was re-run under scikit-learn 1.8.0 with NumPy 2.4.4, which gave identical results to four decimal places. The complete pipeline was implemented in an interactive web application (Sect. 3.8), which reproduced every result reported here end-to-end. Figures were produced with Matplotlib, except Figure 8, which is a screenshot of the application.

11. Use of Generative Artificial Intelligence

Claude (Anthropic) was used to assist with drafting application code and editing the manuscript text. All code was executed, and all outputs, statistics, citations, and text were done by the authors, who take full responsibility for the content.

RESULTS

1. Data Characteristics

Of the 2,190 patients in the raw dataset, 2,188 (99.9%) were retained after auditing (Table 2). The analytic dataset was nearly balanced (1,120 negative, 1,068 positive). The development and held-out sets contained 1,750 (854 positive) and 438 (214 positive) patients, and the 190 flagged records were distributed as 157 and 33, respectively.

TABLE 2: Summary Characteristics of The Analytic Dataset After Data-Quality Auditing.

Characteristic	Value
Total patients (raw dataset)	2,190
Patients retained after quality audit	2,188 (99.9%)
Candidate predictor variables	12 (demographic and haematological)
Malaria-negative, n (%)	1,120 (51.2%)
Malaria-positive, n (%)	1,068 (48.8%)
Missing values; duplicate records	0; 0
Records excluded: WBC > 20,000 cells/μL	2
Records flagged: neutrophil + lymphocyte percentage > 100%	190 (8.7%); retained, assessed in robustness analysis
Development set (80%), n (positive)	1,750 (854)
Held-out test set (20%), n (positive)	438 (214)

2. Minimal-Feature Selection and Cross-Validated Performance

Within the development set, cross-validated AUC varied within a narrow range (0.800–0.813) across feature-set sizes from 12 to 3 (Table 3; Figure 2a). The pre-specified four-feature set achieved the highest mean AUC (0.813 ± 0.025, compared with 0.809 ± 0.031 for all 12 features), and reducing to three features gave a modestly lower value (0.802 ± 0.029); all differences were small relative to fold-to-fold variability. Roughly two-thirds of the candidate features could therefore be removed without loss of discrimination. At k = 4, the same four features (platelet count, RDW-CV, eosinophil count and MCHC) were selected in all five folds.

TABLE 3: Leakage-Safe Five-Fold Cross-Validated Performance Across Feature-Set Sizes in The Development Set (N = 1,750).

Features Retained (k)	AUC (mean ± SD)	Sensitivity	Specificity
12 (all features)	0.809 ± 0.031	0.774	0.685
10	0.802 ± 0.027	0.775	0.678
8	0.803 ± 0.028	0.780	0.683
6	0.800 ± 0.032	0.795	0.671
4	0.813 ± 0.025	0.809	0.676
3	0.802 ± 0.029	0.813	0.673

Feature ranking was recomputed within each training fold, sensitivity and specificity at a threshold of 0.5. AUC: area under the receiver operating characteristic curve; SD: standard deviation across folds.

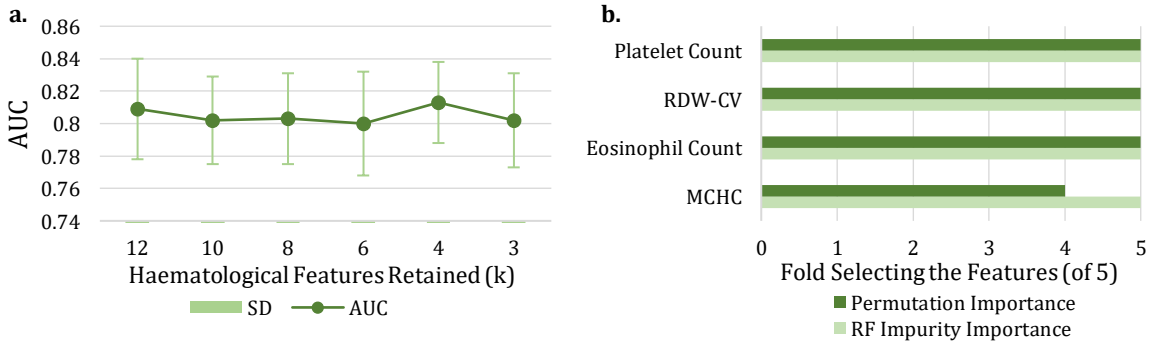


FIGURE 2: Feature reduction and selection stability in the development set. (a) Cross-validated AUC (mean ± SD) by number of retained features; the dotted line marks the 12-feature AUC. (b) Number of folds (of 5) in which each of the four minimal features was selected by permutation importance and RF impurity importance.

3. Cross-checks of The Selected Feature Set

Permutation importance, computed within the same fold structure, selected the identical four-feature set in four of five folds. Platelet count, RDW-CV and eosinophil count were selected in every fold, and in the remaining fold MCHC was replaced by age (Figure 2b). In the development set, Boruta confirmed 10 of 12 candidate features as informative and rejected only haemoglobin and sex (Figure 3). This does not contradict the minimal-feature result: Boruta tests whether a feature carries signal beyond noise, whereas the minimal-feature analysis tests whether a small, jointly sufficient subset reproduces full-panel performance, which is expected when informative haematological indices are mutually correlated.

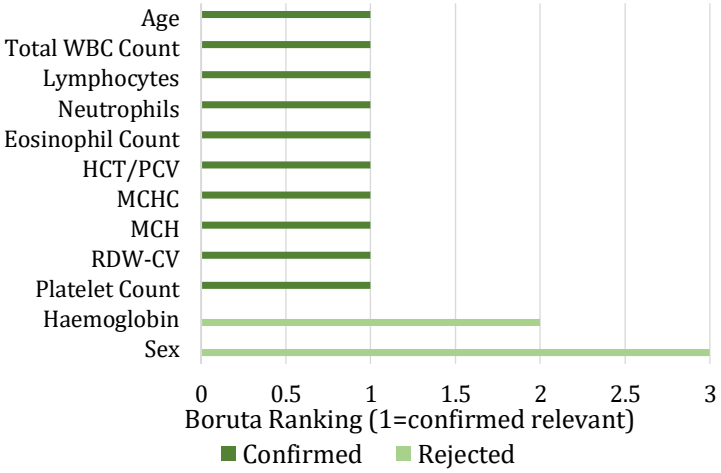


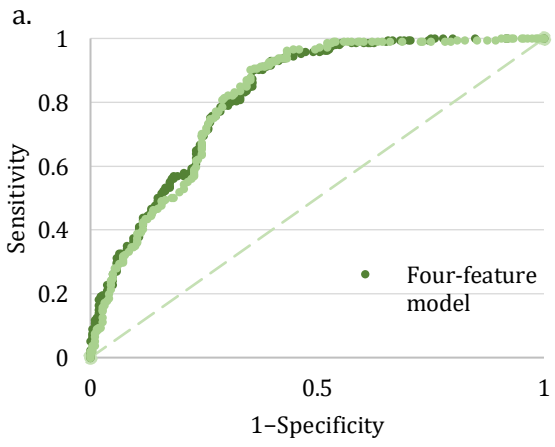
FIGURE 3: Boruta all-relevant feature selection in the development set. Ten of twelve candidate features were confirmed as informative; haemoglobin and sex were rejected as not distinguishable from noise.

4. Held-out Validation and Statistical Comparison

On the held-out test set (n = 438), the four-feature model achieved an AUC of 0.820 (95% CI 0.780–0.856), sensitivity of 0.771 and specificity of 0.714 (Table 4; Figure 4a); at the 0.5 threshold, 165 of 214 malaria-positive and 160 of 224 malaria-negative patients were correctly classified. The point estimate met H3, although the lower confidence limit (0.780) was below 0.80. Calibration was reasonably close to ideal, with mild under-confidence at intermediate probabilities (mean predicted 0.46, observed 0.60) and mild over-confidence at high probabilities (mean predicted 0.89, observed 0.85) (Figure 4b).

TABLE 4: Held-out test-set performance of the four- and twelve-feature random forest models (n = 438; threshold 0.5).

Metric	Four-feature model	Twelve-feature model
AUC (bootstrap 95% CI)	0.820 (0.780–0.856)	0.815 (0.772–0.854)
Sensitivity	0.771	0.813
Specificity	0.714	0.705
Positive predictive value	0.721	0.725
Negative predictive value	0.766	0.798
F1-score	0.745	0.767
Brier score	0.169	0.169
Matthews correlation coefficient	0.486	0.521



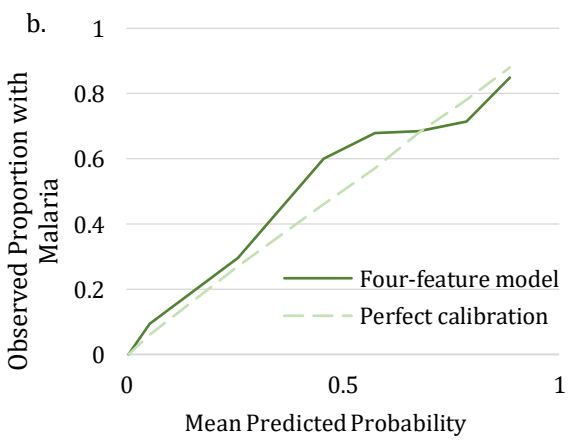


FIGURE 4: Discrimination and calibration on the held-out test set. (a) Receiver operating characteristic curves of the four-feature and twelve-feature models; the dashed diagonal represents chance. (b) Calibration curve of the four-feature model; the dashed line represents perfect calibration.

The AUC of the twelve-feature model on the same patients was 0.815 (95% CI 0.772–0.854). DeLong's test on the paired predictions gave a difference (twelve-feature minus four-feature) of −0.0045 (z = −0.39, p = 0.695), and the 95% CI of this difference (−0.027 to 0.018) lay entirely within the pre-specified margin of ±0.05, supporting H1. The twelve-feature model had slightly higher sensitivity and MCC (Table 4), consistent with its broader information content, but neither difference altered the overall discrimination or clinical utility (Sect. 3.5).

5. Clinical Utility

The four-feature model showed a net benefit greater than both the treat-all and treat-none strategies at every threshold evaluated (0.01–0.89), and the twelve-feature model showed a net benefit between 0.03 and 0.85 (Figure 5; Table 5). The two models' decision curves were close throughout.

TABLE 5: Net benefit at selected threshold probabilities on the held-out test set. The treat-none strategy has a net benefit of 0 at every threshold.

Threshold probability	Four-feature model	Twelve-feature model	Treat-all
0.10	0.450	0.441	0.432
0.20	0.405	0.415	0.361
0.30	0.365	0.375	0.269
0.40	0.307	0.315	0.148
0.50	0.231	0.247	−0.023

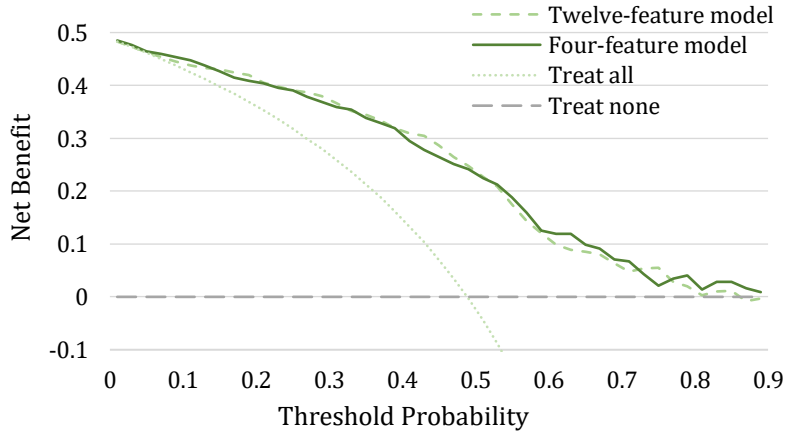


FIGURE 5: Decision curve analysis on the held-out test set: four-feature model (solid line), twelve-feature model (dashed line), treat-all strategy (dotted line), and treat-none strategy (dashed line).

6. Explainability

SHAP ranked platelet count as the dominant contributor (mean |SHAP| = 0.203), followed by RDW-CV (0.071), eosinophil count (0.054), and MCHC (0.051) (Figure 6; Table 6). Lower platelet count (ρ = −0.810), lower RDW-CV (ρ = −0.636), and lower MCHC (ρ = −0.405) were associated with a higher predicted probability of malaria, whereas higher eosinophil count was associated with a higher predicted probability (ρ = +0.304). The dominance of platelet count supports H2.

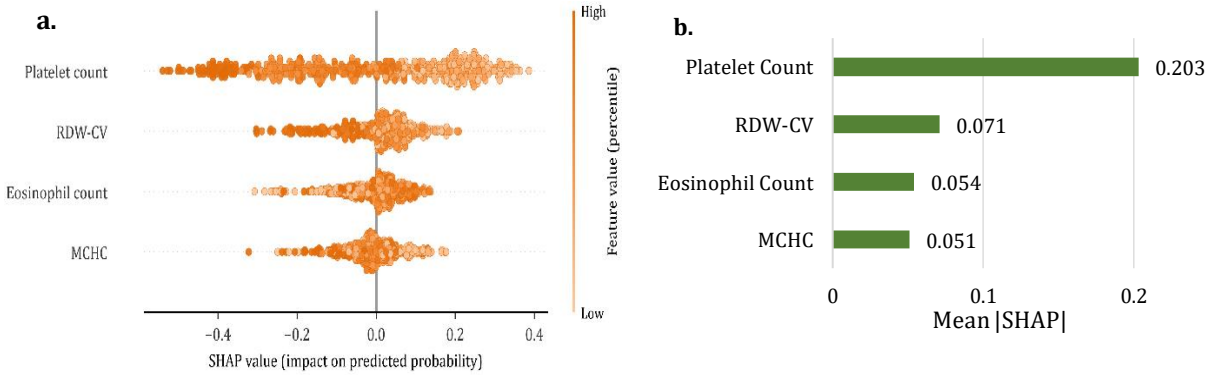


FIGURE 6: SHAP explanations of the four-feature model on the held-out patients. (a) Distribution of SHAP values per patient, shaded from low to high feature value. (b) Global importance as mean absolute SHAP value.

TABLE 6: Global SHAP importance and direction of association for the four selected features.

Feature	Mean |SHAP|	Direction (Spearman ρ)	Interpretation
Platelet count	0.203	−0.810	Lower value → higher predicted probability
RDW-CV	0.071	−0.636	Lower value → higher predicted probability
Eosinophil count	0.054	+0.304	Higher value → higher predicted probability
MCHC	0.051	−0.405	Lower value → higher predicted probability

ρ: Spearman correlation between feature value and SHAP value.

SHAP and LIME identified the same top feature for 71 of 100 sampled patients (71%). Agreement was higher for confidently predicted patients (54/74, 73%) than for patients near the decision boundary (17/26, 65%), consistent with the local linear approximation of LIME being less stable where the decision boundary is complex.

Partial dependence was consistent with SHAP for three features (Figure 7): the average predicted probability fell from 0.72 at the lowest platelet counts to 0.04 above approximately 300 × 10^3/µL, from 0.55 to 0.36 across the range of RDW-CV, and from 0.48 to 0.31 across the range of MCHC. For eosinophil count, the relationship was non-monotonic, peaking at 0.55 around 270 cells/µL and returning to 0.37 at both ends of the observed range. This divergence from the positive SHAP correlation suggests that the contribution of eosinophil count is more interaction-dependent than that of the other features (Sect. 4.4).

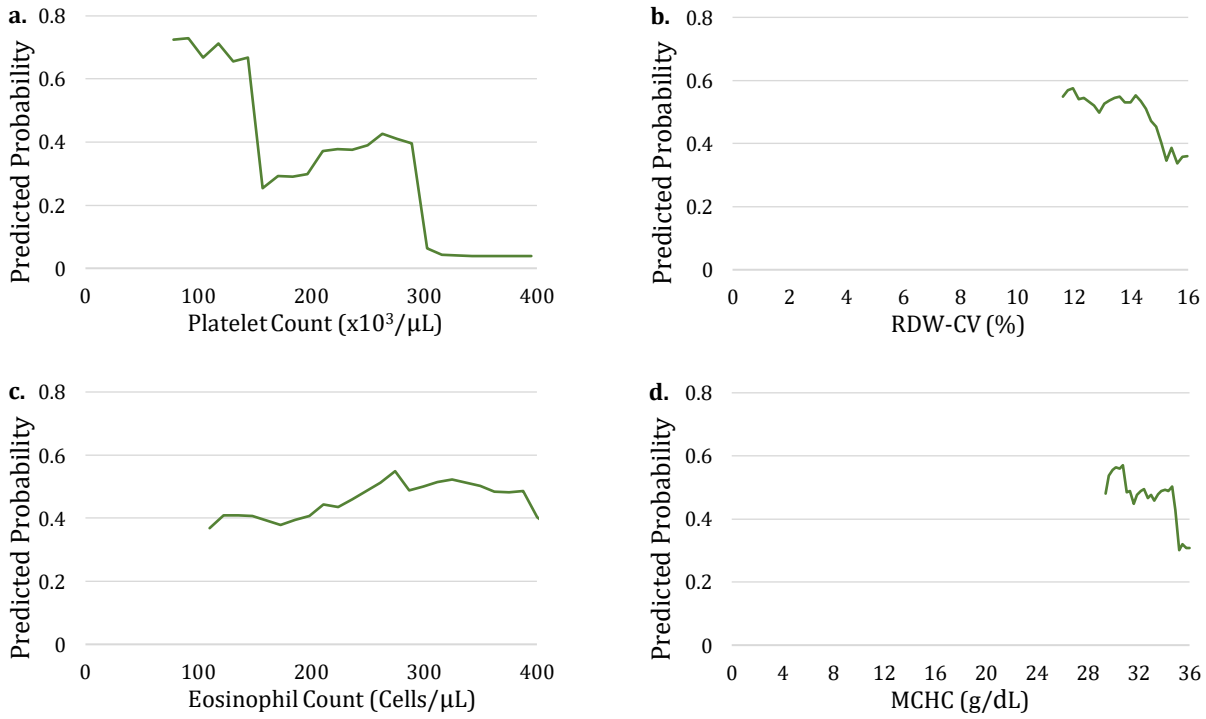


FIGURE 7: Partial dependence of the predicted malaria probability on (a) platelet count, (b) RDW-CV, (c) eosinophil count, and (d) MCHC, computed on the development set.

7. Robustness to The Physiological Data-Quality Flag

In the development set, the cross-validated AUC of the four-feature model was 0.813 with the 157 flagged records included and 0.810 after excluding them (Table 7), a negligible difference consistent with neither neutrophil nor lymphocyte percentage being among the selected features.

TABLE 7: Robustness of the four-feature model's five-fold cross-validated AUC (development set, feature set held fixed) to exclusion of physiologically inconsistent differential-count records.

Scenario	N	Cross-validated AUC
Development set, including flagged records	1,750	0.813
Development set, excluding flagged records	1,593	0.810

8. The PlasmoSense Application

The pipeline was implemented as an interactive web application, PlasmoSense [20], with an English and Indonesian interface selectable by a switch in the header and three modules (Figure 8). The automatic pipeline module runs the full workflow of Sects. 2.2–2.9 in a single action, with a step tracker showing the stages completed (Figure 8a). The manual exploration module lets users choose a feature subset and classifier and run each stage separately for teaching purposes. Its cross-validation is also restricted to the development set, and when the selected subset matches the automatically selected set, it reproduces the results reported here. The patient prediction module returns a predicted probability, the classification at the 0.5 threshold, a qualitative risk category (low < 0.33, moderate 0.33–0.66, high ≥ 0.66) and a patient-specific SHAP explanation, with an on-screen disclaimer that the tool is a research-stage screening aid and not a diagnostic replacement (Figure 8b). Following common practice in prediction-model development, once performance had been estimated on the held-out set, the model used by this module was refitted on all 2,188 patients with the same predictors and hyperparameters. All performance estimates in this article refer to the model trained on the development set. The application does not store uploaded data.

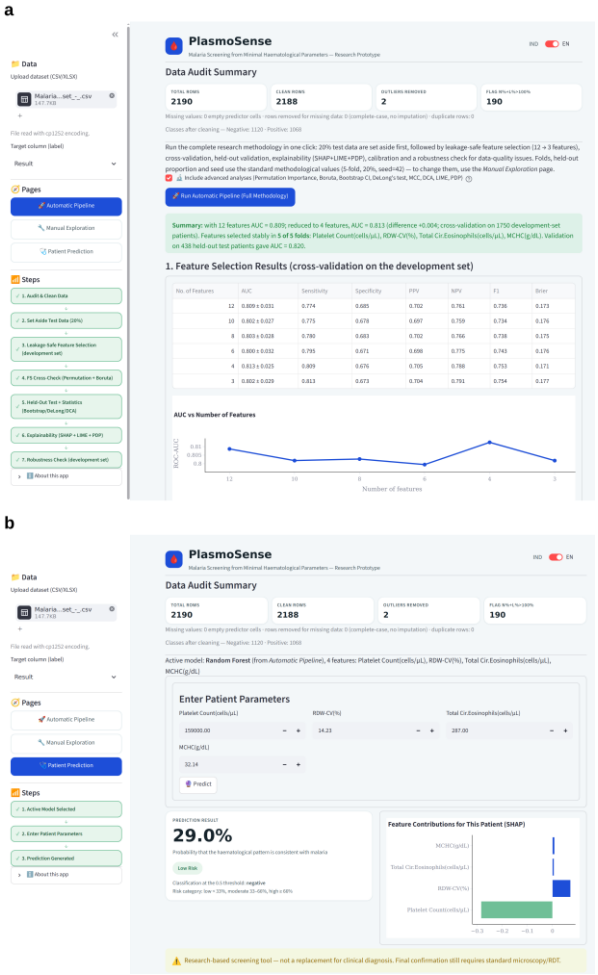


FIGURE 8: The PlasmoSense application (English interface) after analysis of the study dataset. (a) Automatic pipeline module: data-audit summary, completed step tracker (left panel), results summary and the leakage-safe feature-selection results on the development set. (b) Patient prediction module: entered haematological parameters, predicted probability with the classification at the 0.5 threshold and the qualitative risk category, and the patient-specific SHAP explanation.

DISCUSSION

1. Principal Findings

In this secondary analysis of 2,188 Bangladeshi patients, a model restricted to four routine haematological parameters (platelet count, RDW-CV, eosinophil count and MCHC) achieved malaria-screening performance equivalent, within a pre-specified margin, to that of a model using all 12 parameters. This conclusion rests on several consistent findings obtained with a held-out set that played no part in feature selection: similar cross-validated AUCs (0.813 vs 0.809), overlapping bootstrap confidence intervals, a non-significant DeLong's test whose 95% CI for the AUC difference lay within ±0.05, and closely matching decision curves. The minimal set was stable across folds, largely reproduced by permutation importance (identical set in four of five folds, the three leading features in all five) and contextualised by Boruta, which showed that most excluded features carry real but redundant signal. All three pre-specified hypotheses were supported, with the caveat that the lower confidence limit of the held-out AUC (0.780) fell slightly below the H3 threshold.

2. Comparison with Prior Literature

ML, haematology, and explainable artificial intelligence (AI) for malaria classification are individually well precedented. Morang'a et al. [13] found platelet counts among the leading haematological classifiers of uncomplicated malaria in a Ghanaian cohort, which our SHAP results corroborate in a different population. Symptom-based studies from Nigeria have combined ensemble models with SHAP, LIME and permutation importance [2,3], but did not pursue systematic minimal-feature reduction or a formal comparison between reduced and full models. For dengue, Yilmaz et al. [24] applied a leakage-conscious tree-based framework with calibration, decision curve and SHAP analyses to routine haematology from Bangladesh. Our contribution is therefore not ML, explainable AI, or haematology applied to malaria in isolation, but their combination with minimal-feature selection, leakage-safe partitioning and cross-validation, an equivalence-margin comparison, and triangulated explainability on a real, previously unanalysed patient dataset. The exploratory comparators also supported the choice of tree ensembles: Extra Trees performed similarly to RF, whereas logistic regression discriminated less well (held-out AUC 0.740), consistent with non-linear relationships between the haematological predictors and the outcome.

3. Clinical and Translational Implications

A four-parameter model requires no malaria-specific test beyond a standard CBC, which is routinely available in many facilities, including in resource-constrained settings where residual transmission is concentrated (Sect. 1).

We frame the model as a triage aid that flags patients for prioritised confirmatory testing, not as a replacement for parasitological or molecular diagnosis. The DCA result, with positive net benefit across threshold probabilities of 0.01–0.89, offers quantitative support for usefulness across plausible referral thresholds. At the default threshold of 0.5, however, sensitivity was 0.771, so roughly one in four malaria cases would not be flagged. In a triage setting, a lower threshold chosen according to local prevalence and the cost of missed cases would be more appropriate, and the model should not be used to rule out malaria.

4. Explainability Insights and Biological Plausibility

Three of the four findings align with clinical knowledge. The dominant negative association of platelet count matches malaria-associated thrombocytopenia described in clinical studies [13,4], and the negative association of MCHC is compatible with a haemolytic anaemia pattern. The direction for RDW-CV (lower values associated with higher predicted probability) is less intuitive, because haemolysis-related anisocytosis might be expected to raise RDW; we report it as an empirical finding that warrants dedicated investigation rather than as a biological claim. Eosinophil count was the least consistent feature, with a positive SHAP association but a non-monotonic partial dependence. One possible explanation is a helminthic co-infection signal common in malaria-endemic settings, which could not be examined in these data. More broadly, haematology-based malaria detection shows real but heterogeneous diagnostic signal: a meta-analysis of 15 studies reported high pooled accuracy for automated haematology analysers with significant between-study heterogeneity [14], and a West African paediatric study reported that CBC measurements combined with ML classifiers can discriminate malaria from non-malaria cases [17].

5. Strengths and Limitations

The main strengths of this study are methodological: a held-out set removed before any feature selection and fold-wise recomputation of feature rankings, convergent validation of the selected set by two ranking methods, contextualised by Boruta, a formal comparison against a pre-specified equivalence margin rather than point estimates alone, assessment of clinical utility by DCA alongside discrimination and calibration, triangulated explainability with disagreements reported transparently, transparent data auditing with a quantified sensitivity analysis, and full reproducibility across library versions and through a public-facing application.

Several limitations apply. First, all records come from a single country and a secondary, non-adjudicated source. Reference ranges, comorbidities (such as helminthiasis, thalassaemia trait and other haemoglobinopathies) and laboratory methods may differ elsewhere, and the reference-standard test, treatment status and Plasmodium species were not available. Second, there was no external or prospective validation, and performance was estimated on a single held-out partition of moderate size (n = 438), so confidence intervals remain relatively wide. Third, only classical ML with default hyperparameters was used. This was appropriate for the sample size but may not be optimal for larger datasets. Finally, explainability reflects association rather than causation, and the DCA threshold range is illustrative, because the appropriate threshold depends on local prevalence and on the relative costs of missed versus unnecessary confirmatory testing, which were not modelled.

These limitations point to three priorities: external and prospective validation across countries, facility levels and Plasmodium species; dedicated investigation of the eosinophil-count signal, ideally stratified by documented helminthic co-infection; and extension of PlasmoSense into a prospective data-collection and decision-support tool in the resource-limited settings that motivated this work.

CONCLUSION

A random forest model restricted to four routine haematological parameters (platelet count, RDW-CV, eosinophil count and MCHC) achieved malaria-screening performance equivalent, within a pre-specified margin, to that of a twelve-parameter model (held-out AUC 0.820 vs 0.815; 95% CI of the difference −0.027 to 0.018), with comparable clinical utility on decision curve analysis. Explainability analyses converged on platelet count as the dominant and clinically plausible predictor and transparently identified eosinophil count as a feature whose contribution requires further investigation. Because the four parameters are already measured in a standard CBC, this approach may provide a low-cost triage aid for malaria-endemic, resource-limited settings, complementing rather than replacing parasitological and molecular diagnosis, pending external and prospective validation.

STATEMENTS AND DECLARATIONS

Ethics Approval: This study was a secondary analysis of a fully de-identified, publicly available dataset and involved no contact with human participants

Data Availability: The dataset analysed in this study is publicly available in Mendeley Data (Siam et al. 2025): https://doi.org/10.17632/t8n3z7s89h.4.

Code and Application Availability: Available from the corresponding author upon reasonable request.

REFERENCES

[1] Alam, M. S., Alam, M. J., Hossain, M. S., Matin, M. A., Phru, C. S., Hasan, A., Rahman, M. M., Rahaman, M. M., Islam, M. N., Das, S. K., Aktaruzzaman, M. M., Adnan, S. D., Islam, M. N., & Hazarika, A. (2024). Assessing The Role of The For-Profit Private Healthcare Sector in Malaria Elimination Efforts in Bangladesh: A Cross-Sectional Study of Challenges and Opportunities. *BMC Public Health*, 24. https://doi.org/10.1186/s12889-024-20448-2

[2] Attai, K., Ekpenyong, M., Amannah, C., Asuquo, D., Ajuga, P., Obot, O., Johnson, E., John, A., Maduka, O., Akwaowo, C., & Uzoka, F.-M. (2024). Enhancing The Interpretability of Malaria and Typhoid Diagnosis with Explainable AI and Large Language Models. *Tropical Medicine and Infectious Disease*, 9(9), 216. https://doi.org/10.3390/tropicalmed9090216

[3] Awe, O. O., Mwangi, P. N., Goudoungou, S. K., Esho, R. V., & Oyejide, O. S. (2025). Explainable AI for Enhanced Accuracy in Malaria Diagnosis Using Ensemble Machine Learning Models. *BMC Medical Informatics and Decision Making*, 25, 162. https://doi.org/10.1186/s12911-025-02874-3

[4] Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32. https://doi.org/10.1023/A:1010933404324

[5] Bria, Y. P., Yeh, C. H., & Bedingfield, S. (2021). Significant Symptoms and Nonsymptom-Related Factors for Malaria Diagnosis in Endemic Regions of Indonesia. *International Journal of Infectious Diseases*, 103, 194–200. https://doi.org/10.1016/j.ijid.2020.11.177

[6] Chicco, D., & Jurman, G. (2020). The Advantages of The Matthews Correlation Coefficient (MCC) Over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genomics*, 21, 6. https://doi.org/10.1186/s12864-019-6413-7

[7] Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., et al. (2024). TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models that Use Regression or Machine Learning Methods. *BMJ*, 385, e078378. https://doi.org/10.1136/bmj-2023-078378

[8] DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing The Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 44(3), 837–845. https://doi.org/10.2307/2531595

[9] Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection With The Boruta Package. *Journal of Statistical Software*, 36(11), 1–13. https://doi.org/10.18637/jss.v036.i11

[10] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). Curran Associates.

[11] Mariki, M., Mduma, N., & Mkoba, E. (2024). Feature Selection Approach to Improve Malaria Prediction Model's Performance for High- and Low-Endemic Areas of Tanzania. In J. Marx Gómez, A. Elikana Sam, & D. Godfrey Nyambo (Eds.), *Artificial Intelligence Tools and Applications in Embedded and Mobile Systems* (pp. 53–69). Springer. https://doi.org/10.1007/978-3-031-56576-2_6

[12] Minwuyelet, A., Yewhalaw, D., & Atenafu, G. (2026). Diagnostic Challenges in Malaria Detection: A Comparative Diagnostic Performance of HRP2-Based Rapid Diagnostic Tests, Microscopy, and PCR at Bichena Primary Hospital, Northwest Ethiopia. *Parasite Epidemiology and Control*, 33, e00485. https://doi.org/10.1016/j.parepi.2026.e00485

[13] Morang'a, C. M., Amenga-Etego, L., Bah, S. Y., Appiah, V., Amuzu, D. S. Y., Amoako, N., Abugri, J., Oduro, A. R., Cunnington, A. J., Awandare, G. A., & Otto, T. D. (2020). Machine Learning Approaches Classify Clinical Malaria Outcomes Based on Haematological Parameters. *BMC Medicine*, 18, 375. https://doi.org/10.1186/s12916-020-01823-3

[14] Mulatie, Z., Kelem, A., Chane, E., Tarekegn, A. M., Teketelew, B. B., Yenesew, A., et al. (2024). Diagnostic Role of Sysmex Hematology Analyzer in The Detection of Malaria: A Systematic Review and Meta-Analysis. *PLOS ONE*, 19(9), e0296766. https://doi.org/10.1371/journal.pone.0296766

[15] Nsugbe, E., Mathebula, D., Viza, E., Samuel, O. W., Connelly, S., & Mutanga, I. (2023). On the Clinical Use of Artificial Intelligence and Haematological Measurements for a Rapid Diagnosis and Care of Paediatric Malaria Patients in West Africa. *Engineering Proceedings*, 58(1), 131. https://doi.org/10.3390/ecsa-10-16246

[16] Ojurongbe, T. A., Afolabi, H. A., Bashiru, K. A., Sule, W. F., Akinde, S. B., Ojurongbe, O., & Adegoke, N. A. (2023). Prediction of Malaria Positivity Using Patients' Demographic and Environmental Features and Clinical Symptoms to Complement Parasitological Confirmation Before Treatment. *Tropical Diseases, Travel Medicine and Vaccines*, 9, 24. https://doi.org/10.1186/s40794-023-00208-7

[17] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[18] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining The Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

[19] Siam, M. S., Rima, T., Islam, M. R., Syeid, S., & Momin, M. (2025). Malaria-Associated Clinical Data from Bangladesh: A Multivariate Dataset for Epidemiological and Clinical Research (Version 4) [Dataset]. *Mendeley Data*. https://doi.org/10.17632/t8n3z7s89h.4

[20] Streamlit, Inc. (2026). Streamlit (Version 1.64.0) [Computer software]. https://pypi.org/project/streamlit/1.64.0/

[21] Sun, X., & Xu, W. (2014). Fast Implementation of DeLong's Algorithm for Comparing The Areas Under Correlated Receiver Operating Characteristic Curves. *IEEE Signal Processing Letters*, 21(11), 1389–1393. https://doi.org/10.1109/LSP.2014.2337313

[22] Vickers, A. J., & Elkin, E. B. (2006). Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making,* 26(6), 565–574. https://doi.org/10.1177/0272989X06295361

[23] World Health Organization (2025). World Malaria Report 2025. *World Health Organization*, Geneva.

[24] Yilmaz, Z. B., Kucukakcali, Z., & Akbulut, S. (2026). Tree-based Machine Learning for Diagnostic Classification of Dengue Fever Using Routine Hematological Parameters: A Secondary Analysis of a Publicly Available Dataset. *Diagnostics*, 16(18), 2966. https://doi.org/10.3390/diagnostics16182966