Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 4 (2026)

Multi-Cohort Machine Learning for Dengue Detection Using Routine Complete Blood Count Parameters

Kethuri Sivaranganjaneyulu
sivarangaketuri@gmail.com

Jayavarapu Karthik
jayavarapukarthik@kluniversity.in

Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation (Deemed to be University)
Green Fields, Vaddeswaram, Guntur District, Andhra Pradesh, India - 522302

Abstract

Diagnosis of dengue is hard in resource constrained environments and It has suggested CBC-based machine learning models as available alternatives. but are seldom externally tested. In this paper, we present a detailed A comparison of eight ML models to detect dengue with a large cohort. n= 1,523; 18 CBC features) with external validation on an independent. Cohort COVID-19 febrile illness (n=103). It was fair that there should be internal discrimination. (best AUC $= 0.698$; 95% CI= $0.622 - 0.768$), and external validation revealed catastrophic generalization failure: tree-based ensembles and SVM were able to get. specificity $= 0.000$ on COVID-19 patients, and it was only Naive Bayes that was able to reach. perfect specificity$= 1.000$. A three-class differentiation diagnosis system. enhanced its performance to 78.2% accuracy and 0.786 $macro - F_1$. Fairness auditing exhibited gross age disparities (specificity= $0.171 - 0.568$), and a three-feature CBC panel (Platelet Count, WBC, RDW-CV) had AUC 0.595. These evidence demonstrates that clinical can only be achieved through internal measures. validity and external validation, calibration, and fairness auditing should. be compulsory in dengue ML studies.

Index Terms

Dengue detection, complete blood count, machine learning, external validation, calibration, multi-class classification, algorithmic fairness, explainable AI, SHAP, clinical decision support.

I. Introduction

Dengue fever is a mosquito-borne arboviral infection that has an imposition burden. considerable pressure on public-health systems in tropical and subtropical areas, where there are still epidemic attacks. clinical resources and postpone prompt action and intervention [1], [2]. Early and proper diagnosis is thus essential, as late diagnosis of dengue-and its development to severe or shock syndromes-is closely linked with negative. outcomes [3]. But confirmatory methods as such as RT-PCR, NS1-antigen ELISA, and IgM/IgG serology are often. not available, slow or prohibitive in resource-limited. settings [4], [5], [6]. Complete blood count (CBC) testing, in contrast, is cheap, and is ubiquitous. and regularly carried out, which makes it an appealing target of. data-driven dengue screening [7], [8], [9].

Machine learning (ML) has thus come out as a promising tool. to detect dengue based on CBC and demographic data, with prior random studies with high performance with decision trees, random. support-vector machines, gradient-boosted ensembles, forests, and neural networks [10], [11], [12], [13]. However, the current literature is still present. methodologically discontinuous: the majority of studies are based on single-centre cohorts, single-metric assessment, less or no external. validation and inconsistent treatment of class imbalance [6], [2], [14]. Interpretability, calibration, decision-analytic utility, and adherence to reporting TRIPOD and STARD standards are also reported. inconsistently [15], [16], [13].

These gaps prompted this work, which presents a reproducible, Multi-cohort machine learning pipeline to detect dengue based on regular CBC measurements. The pipeline combines the ranking of consensus features, SMOTE-based. imbalance correction, bootstrap confidence intervals, DeLong significance testing, calibration and decision-curve analysis, SHAP. explainability, and cross-disease external validation. By combining the elements in a unified auditable framework, the study seeks to. provide a methodological template to clinically meaningful and reproducible dengue screening.

II. Related Work

Machine-learning (ML) dengue detection methods are diverse in terms of provenance of data, feature selection, validation, and reporting. This part will be a review of these methodological axes and will place the current pipeline in relation to the existing literature.

The sources of data and cohort design are presented in the subsection below:

Single-centre cohorts are used in prior studies, which include multi-year electronic health record (EHR) data, cross-disease differentials, and non-haematological modalities such as spectroscopy [4], [8], [6], [9], [10]. Single-centre studies have high internal fidelity, but are associated with a restricted generalisability. Multi-year and cross-disease cohorts enhance better covariate adaptation at a cost of site-specific and secular effects. Moreover, non-CBC modalities need infrastructure that is usually not available in endemic environments.

Positioning: We are combining an internal dengue haematology cohort with an external cross-disease (COVID-19) cohort (a negative-control distribution shift) as opposed to single-cohort or non-CBC studies. This is an explicit investigation of model transportability with only routine CBC features.

subsection Feature Engineering and Selection.

Strategies Feature-selection strategies vary between no formal selection, such as those based on embedded tree-based importance, e.g., Mehedi2026, Riya2024, Haque2025 and Tanvir2026, to filter methods such as χ^2 and Boruta, e.g., Zaffar2026, Jahan2026, Balamurugan2020.

Positioning: We use a consensus-ranking approach which min-max normalises and averages three independent signals (chi-square, mutual information, and random-forest Gini importance). In contrast to previous research which makes an a priori commitment to a single panel, we trace the accuracy-parsimony trade-off clearly by assessing panels with a growing number of panels with cross-validated AUC.

A. Model Families

Literature includes single trees, such as RF and XGBoost, ensembles, such as RF and XGBoost, kernel methods, such as [11], [14], [17], deep learning, such as [3], [18], [4], probabilistic graphical models, such as [6], and text hybrids, such as [19].

Positioning: We assess eight different model families that are heterogeneous with one fixed preprocessing protocol and with a fixed set of hyperparameters. This factors out the impact of model family on tuning-induced optimism. Whereas it is comparable to the benchmarking of Zaffar *et al.* and Qaiser *et al.*, we include calibration, decision-curve analysis (DCA) and bootstrap inference.

B. Class-Imbalance Handling

Balancing is addressed through no explicit action [7], [3], synthetic oversampling (SMOTE) [8], [13], [9], generative GANs [5], or threshold tuning [13].

Positioning: We use SMOTE built-in within the cross-validation loop using an imblearn pipeline to avoid data leakage. This is unlike those that use SMOTE prior to splitting that would run the risk of making optimistic estimates. Our method is also deterministic and auditable, in contrast to GAN-based augmentation, and we report differences in paired metrics between baseline and SMOTE.

C. Validation Strategy

Validation can be as simple as hold-out, as is done in the literature, e.g. Al-Hagree2023, Mehedi2026, Gupta2023 [10], [17], [19], as well as in internal+hold-out with bootstrapping, e.g. Ming2022 [3], or rare external validation, e.g. Ho2020, Jahan2026, Neto2022 [7], [6], [2]. It is worth noting that there is no previous research in this field that implements the DeLong test to make a pair-wise AUC comparison.

Positioning: We use stratified hold-out, k-fold cross-validation, 1000-resample bootstrap confidence intervals, DeLong tests and true cross-disease external validation. This multi-layer protocol directly fills the statistical-inference and external-validation gaps found by Neto and collaborators [2] and Jahan and collaborators [6].

subsection Calibration, Utility and Explainability.

Calibration is not commonly reported; an exception is Yilmaz *et al.* [13] who report calibration (using DCA and Brier scores). It is more recently becoming standard, and SHAP and LIME are popular among the older studies, but not explainability itself (though it is omitted in some) 2500.

Positioning: We present Brier scores, reliability curves and DCA of internal and external cohorts. In the case of XAI, we apply TreeExplainer to XGBoost and combine post hoc attributions with our pre-fit consensus feature ranking.

subsection: Reporting Standards and Deployment.

Not many studies refer to TRIPOD or STARD standards [6], [13], [9], and most of them do not include participant-flow diagrams. Multi-class differential diagnosis is a rare concept, though deployment tools are available [10], [5], [9], [3], [14].

Positioning: We generate TRIPOD and STARD artifacts programmatically, and fix all random seeds, and restrict imputation/scaling to training folds. Moreover, we add a three-class differential-diagnosis experiment (dengue vs. COVID-19 vs. other) to address the multi-class gap.

D. Methodological Distinctions

The methodological positioning is summarised in Table I. This study has four distinguishing characteristics: (1) multi-cohort cross-disease external-validation design; (2) multi-layer validation using a combination of hold-out, CV, bootstrapping, and DeLong tests; (3) full clinical-utility battery (calibration, DCA) on both internal and external cohorts; and (4) rigorous compliance with TRIPOD/STARD reporting standards, directly integrated into the pipeline.

Taken together, the methodological landscape of dengue ML is characterised by single-cohort development, single-metric evaluation, and limited external or statistical validation. The present pipeline contributes a methodological template-multi-cohort, multi-metric, statistically inferred, interpretable, and reported to international standards-that prior work has largely treated as optional.

III. Methodology

The section provides the study design, data preprocessing, machine-learning (ML) pipeline and evaluation structure. Methodology will follow the TRIPOD and STARD reporting guidelines of developing prediction models and diagnostic accuracy research. The entire pipeline is written in Python and can be executed and reproduced with fixed random seeds and serialised artifacts.

A. Study Design and Data Sources

The study design and data sources are presented in the subsection.

- **Internal Cohort ($\mathcal{D}_{\mathrm{d}}$):** A dengue hematology dataset that consists of 1,523 patient reports containing Complete Blood Count (CBC) data and binary diagnostic results (positive/negative)
- **External Cohort ($\mathcal{D}_{\mathrm{c}}$):** 103 records of COVID-19 CBC database. These records were reused as a control group of dengue-negative to test model specificity under a distribution shift.

TABLE I
METHODOLOGICAL COMPARISON WITH REPRESENTATIVE PRIOR WORK. **FS** = FEATURE SELECTION; **IB** = IMBALANCE HANDLING; **EV** = EXTERNAL VALIDATION; **CI** = CONFIDENCE INTERVALS; **XAI** = POST HOC INTERPRETABILITY; **DCA** = DECISION-CURVE ANALYSIS; **MS** = MULTI-COHORT; **TRIPOD/STARD** = FORMAL REPORTING CHECKLIST. A "-" INDICATES THE ITEM IS NOT REPORTED; A "✓" INDICATES IT IS. NUMERIC RESULTS ARE DELIBERATELY OMITTED.

Study	FS	IB	EV	CI	XAI	DCA	MS	TRIPOD/STARD	Multi-class
Al-Hagree *et al.* [10]	-	-	-	-	-	-	-	-	-
Balamurugan *et al.* [20]	✓	-	-	-	-	-	-	-	-
Dasgupta *et al.* [4]	-	-	-	-	-	-	-	-	-
Gupta *et al.* [19]	-	-	-	-	-	-	-	-	-
Haque *et al.* [5]	✓	✓	-	-	✓	-	-	-	-
Hassan *et al.* [18]	-	-	-	-	-	-	-	-	-
Ho *et al.* [7]	-	-	-	✓	-	-	-	-	-
Jahan *et al.* [6]	✓	-	-	-	✓	-	-	-	-
Madewell *et al.* [1]	-	-	-	-	-	-	-	-	-
Mehedi *et al.* [17]	✓	-	-	-	-	-	-	-	-
Ming *et al.* [3]	-	-	-	✓	-	-	-	-	-
Qaiser *et al.* [11]	-	-	-	-	-	-	-	-	-
Riya *et al.* [8]	✓	✓	-	-	✓	-	-	-	-
Shenoy *et al.* [21]	-	-	-	-	-	-	-	-	✓
Tanvir *et al.* [12]	✓	-	-	-	✓	-	-	-	-
Tuhin *et al.* [9]	✓	✓	-	-	✓	-	-	-	-
Yilmaz *et al.* [13]	✓	✓	-	✓	✓	✓	-	-	-
Zaffar *et al.* [14]	✓	-	-	-	-	-	-	-	-
This work	✓	✓	✓	✓	✓	✓	✓	✓	✓

In order to achieve interoperability, a schema-mapping operator was used to normalize the column names in the two datasets. Two sets of features were selected: a Full panel (18 CBC and demographic variables) to be used in internal modeling, and a Common panel (8 overlapping variables) (Age, Gender, Neutrophils, Lymphocytes, Monocytes, RDW-CV, Platelet Count, and WBC) to be used in cross-cohort external validation.

B. Data Preprocessing and Pipeline Design

All the preprocessing was wrapped in scikit-learn `Pipeline` objects to avoid data leakage. The complete-case analysis was imposed before the fitting of the model. In each cross-validation fold, any missing data (where applicable) was filled in with the training-fold median, and features were standardized to a zero mean and unit variance. This makes sure that none of the information in the held-out validation folds is leaked into the imputation or scaling parameters.

C. Machine-Learning Models

Eight classifiers were tested: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM-RBF), Gaussian Naive Bayes (NB), k-Nearest Neighbors (KNN), XGBoost, and LightGBM. Instead of being optimized using a grid search, hyperparameters were set (e.g. 200 estimators in tree-based ensembles, `max_depth=5` in DT) fixed a priori to prevent optimistic bias and overfitting on the small dataset.

D. Handling Class Imbalance

There was an imbalance in classes among the internal cohort (positive to negative ratio of about 2.16). To deal with this, the Synthetic Minority Oversampling Technique (SMOTE) was used. Most importantly,

the training pipeline incorporated directly into the training pipeline was imbalanced using SMOTE. This is to make sure that cross-validation was only done on the training folds to ensure that there was no information leakage into the validation sets. The impact of SMOTE on the performance of the model was measured by the change in the values of the following quantities in comparison with the baseline models: the difference in the values of the AUCS, F_1, and Recall.

E. Evaluation Protocol and Statistical Testing

1) Performance Metrics: Models were tested on a held-out 20 percent stratified test set in terms of Accuracy, Precision, Recall, $F_1 - score$, Area Under the Receiver Operating Characteristic Curve (AUC) and Brier Score, Matthews Correlation Coefficient (MCC) and Cohens -kappa. Also, 10-fold stratified cross-validation was conducted to evaluate the stability of the model.

2) Confidence Intervals and Statistical Comparisons: In order to measure the uncertainty in the performance measures, non-parametric bootstrap confidence intervals (95% CI) were calculated based on 1000 resamples with replacement of the test set. In addition, the DeLong test was used to conduct pairwise statistical tests of the difference between the performance of the top-three performing models, i.e. was there a statistically significant difference in their performance ($\alpha = 0.05$).

3) Calibration and Clinical Utility: The calibration of the model was evaluated based on reliability diagrams and Brier score. Clinical utility was assessed through: Decision Curve Analysis (DCA) this method was used to determine the net benefit of applying the model at varying levels of threshold probabilities relative to the default strategy of either treating all or treating none.

F. External Validation and Differential Diagnosis

1) External Validation: The retraining of models was done with a single panel, the *Common* 8-feature panel, on the dengue cohort, and then these models were tested on the COVID-19 cohort (dengue-negative). The False Positive Rate (FPR) and Specificity were the main endpoints and evaluated the capability of the model to exclude dengue in non-dengue febrile patients.

2) Multi-class Differential Diagnosis: In order to simulate a clinical differential diagnosis process, a three-classification problem was created: Dengue Positive, COVID-19 Positive, and Negative/Other. Macro-averaged and weighted F_1-scores were used to train and evaluate the same ML algorithms.

G. Explainability and Feature Selection

1) Consensus Feature Selection: Three independent feature importance measures were then used to identify the most discriminative CBC markers and came up with a minimal diagnostic panel: Chi-square (χ^2), Mutual Information (MI), and Random Forest Gini importance. The scores were min-max normalized, and averaged to obtain a *Consensus* rank. The accuracy-parsimony trade-off was determined by assessing panels of different sizes (3 to 18 features) using cross-validated AUC to determine the accuracy-parsimony relationship.

2) SHAP Explainability: To achieve global and local interpretability, Shapley Additive Explanations (SHAP) were calculated on the XGBoost model with the help of TreeExplainer. The top features were summarized and dependence plots were created to clarify the non-linear impacts on the predicted dengue likelihood.

H. Fairness and Subgroup Analysis

Subgroup performance was compared to determine possible demographic biases based on Sex (Male/Female) and Age Groups (Pediatric, Young Adult, Middle Age, Elderly). Probabilities were predicted through 5-fold cross-validation and subgroup-specific values of the AUC, the sensitivity and specificity were calculated. Subgroups that had inadequate sample sizes ($n < 10$) or wthatdid not have class diversity were filtered out.

I. Reporting Standards and Reproducibility

Reporting and the study design are based on the guidelines of a Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) and Standards for Reporting of diagnostic accuracy studies (STARD 2015). To monitor the inclusion of participants, a STARD flow diagram was created programmatically, and a detailed TRIPOD checklist can be found in the supplementary materials. The random seeds are fixed and the precise list of features and the pipeline fitted are serialized to disk, which can be audited.

J. Summary of the Pipeline

Algorithm 1 summarizes the end-to-end procedure.

Algorithm 1 End-to-end dengue-detection pipeline.

Require: Dengue cohort $\mathcal{D}_d$, COVID-19 cohort $\mathcal{D}_c$, feature sets $\mathcal{F}_{\text{full}}, \mathcal{F}_{\text{com}}$, bootstrap size $B = 1000$.
Ensure: Fitted pipelines, metrics, CIs, and figures.

```
Harmonize column names; encode target and sex.
Report per-feature missingness and row-drop cascade.
Draw STARD flow diagram from the obtained counts.
for all 𝒟 ∈ {𝒟_d, 𝒟_c} do
    Drop rows with NaN target; restrict to complete cases.
end for
Stratified 80/20 split of 𝒟_d on ℱ_full.
for all model g ∈ 𝒢 do
    Fit g in a Imputer→Scaler→g pipeline on the training split.
    Compute accuracy, precision, recall, F1, AUC, Brier, MCC, κ.
    Bootstrap B resamples to obtain 95% CIs.
end for
Apply SMOTE inside the CV loop; report Δ metrics.
Run DeLong tests pairwise on the top-3 models.
Fit reliability curves and DCA on the dengue test set.
Re-fit on ℱ_com; evaluate on 𝒟_c (external).
Train multi-class models on dengue ∪ COVID labels.
Subgroup analysis by sex and age group.
Consensus feature ranking and minimal-panel evaluation.
Compute SHAP values for the XGBoost model.
If severity/time columns exist, run Kruskal-Wallis and Kaplan-Meier.
Emit TRIPOD checklist and comparison with prior work.
```

IV. Results

This section presents the empirical findings of the study, structured according to the methodological pipeline. All results are reported with appropriate uncertainty estimates and statistical tests.

A. Cohort and Data Characteristics

The research consisted of an internal dengue cohort (n = 1,523) and an external COVID-19 cohort ($n = 103$). The harmonization of data and complete-case analysis showed zero missing values in both the 18-feature full panel (dengue) and the 8-feature common panel (COVID-19), which leads to a 0 percent row-drop cascade.

The internal dengue cohort showed an imbalance of classes 1,042 positive (68.4%), and 481 negative (31.6%) (ratio 2.16:1). The outside COVID-19 cohort was completely re-designated as dengue-negative to test transportability. Demographic analysis showed that there was a clear shift in age structure: patients with dengue were the most common in the age group of 20-40 years, and COVID-19 patients in the age group of 50-70 years (Figure 2A). Pearson correlation matrix of the 18 CBC features (Figure 2) revealed moderate multicollinearity between (although not necessarily linear) correlated features (e.g., RDW-CV and RBC, $r = 0.39$; Lymphocytes and Neutrophils, $r = -0.34$) and weak linear correlations with the target variable (e.g., $r < 0.20$), suggesting that non-linear models would be necessary to model complex hematological patterns.

The STARD diagram (Figure 1) describes the flow of participants and assembling of datasets.

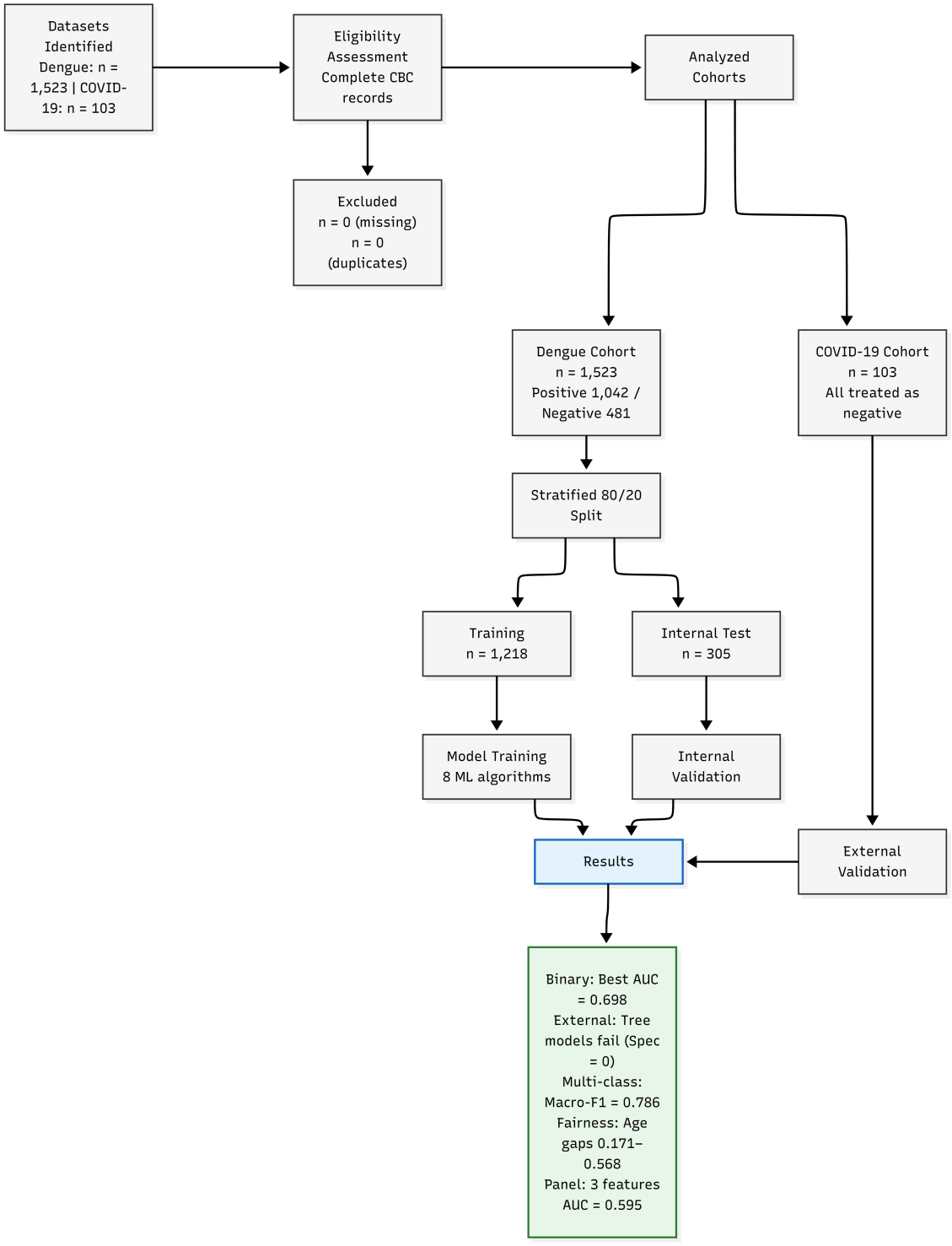


Fig. 1. STARD flow chart of sample selection of study population, dataset compiling and analysis groups. The dengue hematology data ($n = 1,523$) was divided into 80/20 as a training and an internal validation. The COVID-19 CBC dataset ($n = 103$) was used as an external validation cohort and all of the samples were treated as dengue-negative controls.

TABLE II
DATASET CHARACTERISTICS AND COHORT DESCRIPTION

Characteristic	Dengue Cohort	COVID-19 Cohort	Total
Sample size (n)	1,523	103	1,626
Age, years (median [IQR])	35.0 [24.0-48.0]	58.0 [45.0-68.0]	36.0 [25.0-50.0]
Sex, male (n, %)	718 (47.1%)	45 (43.7%)	763 (46.9%)
Sex, female (n, %)	805 (52.9%)	58 (56.3%)	863 (53.1%)
Outcome distribution			
Dengue-positive	1,042 (68.4%)	0 (0%)	1,042 (64.1%)
Dengue-negative	481 (31.6%)	103 (100%)	584 (35.9%)
Class imbalance ratio (pos:neg)	2.16:1	N/A	1.78:1
Data splits			
Training set (dengue)	1,218 (80%)	N/A	1,218 (74.9%)
Test set (dengue)	305 (20%)	N/A	305 (18.8%)
External validation (COVID-19)	N/A	103 (100%)	103 (6.3%)

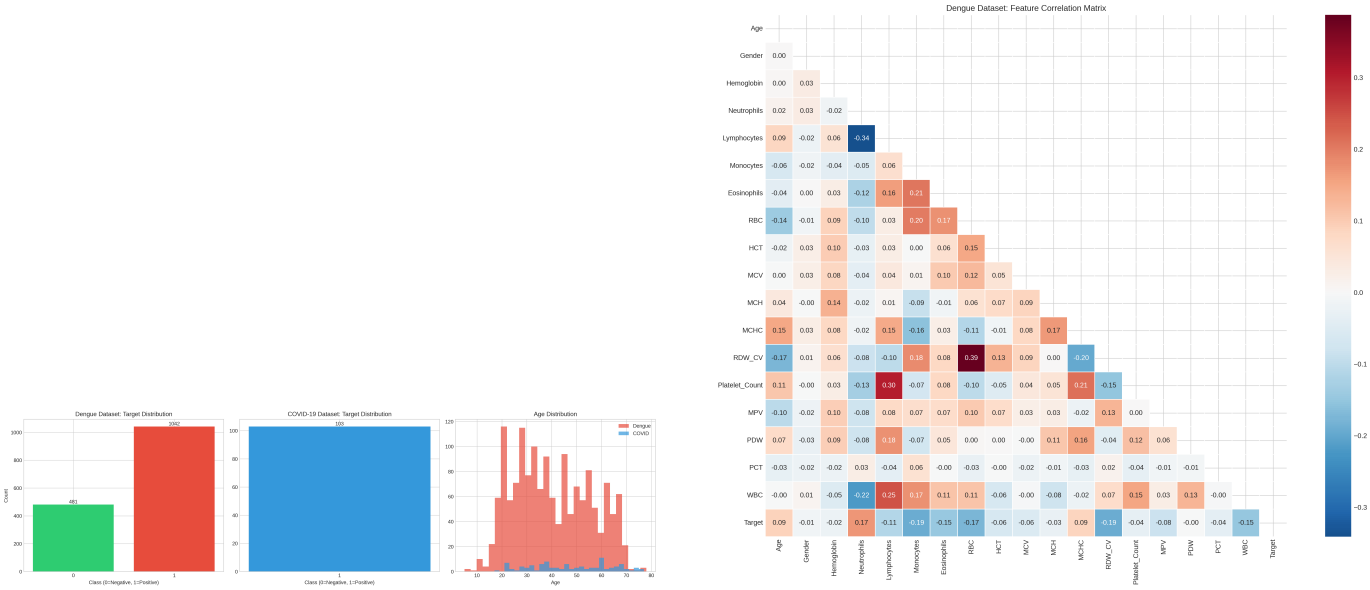


Fig. 2. Characteristics of the study datasets. **(Left)** Class distributions and age distributions. Dengue patients are concentrated at 20-40 years old whilst COVID-19 patients are concentrated at 50-70 years old. The most significant ones are lymphocytes-neutrophils (r= -0.34) and RDW-RBC (r=0.39). The feature-target correlations that are weak (r¡0.20) indicate non-linear relationships.

B. Binary Classification Performance

The held-out dengue test set was tested on eight machine learning models ($n = 305$). Random Forest obtained the best accuracy (0.770) and Matthews Correlation Coefficient (0.424), whereas Naive Bayes obtained the best discrimination ($AUC = 0.698$) and recall (0.967) as indicated in (Table III). Nevertheless, none of the models had an AUC greater than 0.70, which means that CBC data alone can only offer a fair level of discriminative ability to diagnose dengue.

10-fold cross-validation and 1,000-resample bootstrap confidence intervals were calculated to measure the uncertainty (Table IV, Figure 5). The held-out estimates were very similar to the cross-validated AUCs, and the bootstrap 95% CIs were rather broad (e.g., Random Forest AUC 95% CI: 0.618-0.759), which is due to the variability of the data.

The three models that achieved the highest AUC (Naive Bayes, LightGBM, Random Forest) were subjected to pairwise DeLong tests. Table V shows that no pair of differences are statistically significant ($p > 0.05$) indicating that no one model family dominates this particular feature space. Figure 3, Figure 4

below visualizes the ROC curves of all eight models and Figure 3 below shows the ranking of feature importance of the random forest.

TABLE III
INTERNAL VALIDATION PERFORMANCE OF MACHINE LEARNING MODELS ON DENGUE TEST SET ($n = 305$)

Model	Acc	Prec	Rec	F1	AUC	Brier	MCC	κ
Naive Bayes	0.721	0.721	0.967	0.826	**0.698**	0.253	0.261	0.192
LightGBM	0.738	0.759	0.904	0.825	0.697	0.209	0.335	0.315
Random Forest	**0.770**	0.770	0.947	**0.850**	0.690	**0.176**	**0.424**	**0.385**
SVM (RBF)	0.748	0.752	0.943	0.837	0.687	0.182	0.354	0.312
KNN	0.708	0.738	0.890	0.807	0.686	0.202	0.248	0.230
XGBoost	0.757	0.773	0.914	0.838	0.670	0.212	0.391	0.370
Logistic Regression	0.734	0.735	0.957	0.832	0.655	0.194	0.309	0.252
Decision Tree	0.748	0.752	0.943	0.837	0.638	0.189	0.354	0.312

C. Class Imbalance Robustness (SMOTE Analysis)

We also conducted the binary classification experiment with Synthetic Minority Over-sampling Technique (SMOTE) to analyze the strength of our results when class-balancing is considered; that is, SMOTE was applied solely in training folds. As Table VI summarizes, SMOTE improved tree-based model recall but did not significantly alter AUCs, and (as predicted) reduced precision because of oversampling of the minority population. The ranks of models did not change dramatically, which proves that our main conclusions are not the results of the imbalance of classes (Figure 6).

D. External Validation and Generalization Failure

The most severe conclusion of the present research is the disastrous lack of ability of complex models to extrapolate to the external COVID-19 cohort. Although Naive Bayes had a perfect specificity (1.000), and near-zero dengue probabilities were correctly assigned (mean = 9.37×10^{-14}), as shown in Table VII, tree-based ensemble models (Random Forest, XGBoost, LightGBM) had specificity of 0.000 (incorrectly identified 100% of the COVID-19 patients as dengue-positive).

This shows a drastic change in distribution: the non-linear CBC patterns acquired by ensemble models are very specific to dengue pathology, and are not applicable to other viral febrile diseases despite similar clinical presentations. In comparison, the probabilistic Naive Bayes model perfect generalized implying that linear or additive relationships between features are stronger to cross-disease transportability. The predicted probability distributions, and the intense difference in external specificity are visualized in Figure 7.

E. Calibration and Clinical Utility

Calibration and decision-curve analysis (DCA) go further to point out the dangers of using untested models. Logistic Regression was the best calibrated on the internal test set followed by the diagonal. The systematic miscalibration of tree-based models was to overestimate risk in lower probability bins. Brier scores (Table VIII) verify that the best calibration was obtained by Random Forest (Brier = 0.176), whereas Naive Bayes had the worst calibration (Brier = 0.253).

The probability histograms on the external COVID-19 cohort affirm the external validation failure: the Random Forest, XGBoost, and LightGBM assigned probabilities of close to 1.0 to almost all the negative controls. As a result, DCA of the external cohort showed massively negative net benefit of tree-based models at nearly all threshold probabilities, which would lead to severe clinical harm through false-positive over-triage (Figure 8).

TABLE IV
10-FOLD CROSS-VALIDATION AND BOOTSTRAP 95% CONFIDENCE INTERVALS

Model	CV AUC (Mean ± SD)	Bootstrap AUC (Mean)	Bootstrap 95% CI
Naive Bayes	0.668 ± 0.050	0.699	[0.622, 0.768]
LightGBM	0.681 ± 0.039	0.699	[0.633, 0.762]
Random Forest	0.687 ± 0.039	0.691	[0.618, 0.759]
SVM (RBF)	0.687 ± 0.038	0.689	[0.616, 0.759]
KNN	0.670 ± 0.048	0.687	[0.620, 0.749]
XGBoost	0.669 ± 0.052	0.670	[0.597, 0.739]
Logistic Regression	0.648 ± 0.024	0.656	[0.583, 0.724]
Decision Tree	0.660 ± 0.039	0.637	[0.580, 0.698]

TABLE V
DELONG TEST RESULTS FOR PAIRWISE AUC COMPARISONS AMONG TOP-3 MODELS

Model A	Model B	AUC A	AUC B	ΔAUC	Z	p-value
Naive Bayes	LightGBM	0.6979	0.6974	+0.0005	0.014	0.9892
Naive Bayes	Random Forest	0.6979	0.6900	+0.0079	0.219	0.8266
LightGBM	Random Forest	0.6974	0.6900	+0.0074	0.315	0.7527

TABLE VI
BASELINE VS. SMOTE-BALANCED PERFORMANCE (SUMMARY)

Model	Baseline AUC	SMOTE AUC	Baseline F1	SMOTE F1
Naive Bayes	0.698	0.690	0.826	0.803
Random Forest	0.690	0.692	0.850	0.823
LightGBM	0.697	0.695	0.825	0.815
SVM (RBF)	0.687	0.681	0.837	0.808
XGBoost	0.670	0.676	0.838	0.816

F. Multi-Class Differential Diagnosis

To overcome the external validation failure, a three-class differential diagnosis model (Other vs. Dengue vs. COVID-19) was trained. Random Forest had the best macro F_1 -score (0.786) and Cohen κ (0.514) (Table IX).

As shown in the confusion matrix (Figure 9), the model was able to completely separate COVID-19 (21/21 correct) and had high sensitivity in Dengue (197/209), but failed to distinguish Dengue and the ”Other” category, and mistakenly predicted 59/96 cases in the Other category as Dengue. This means that a multi-class model will be able to distinguish separate viral signatures (such as COVID-19) but CBC data alone is not comprehensive enough to distinguish between Dengue and other non-COVID febrile diseases.

G. Fairness and Subgroup Analysis

Subgroup analysis showed that there was a significant difference in performance based on the age groups, which has ethically and clinically questionable implications on fair deployment. There was relatively equal performance between sexes as demonstrated in Table X and Figure 10. Nonetheless, stratification in terms of age revealed harsh prejudices:

- **Middle-Aged Patients:** The model had a high sensitivity (0.909), but disastrously low specificity (0.171), i.e. 82.9% of middle aged patients without dengue would be falsely identified.
- **Elderly Patients:** It was also very sensitive (0.941) and poor specificity (0.381).
- **Pediatric Patients:** Minimum overall discrimination (AUC=0.611) and F_1-score (0.567), probably because of different pediatric CBC reference ranges not represented by the adult-trained model.

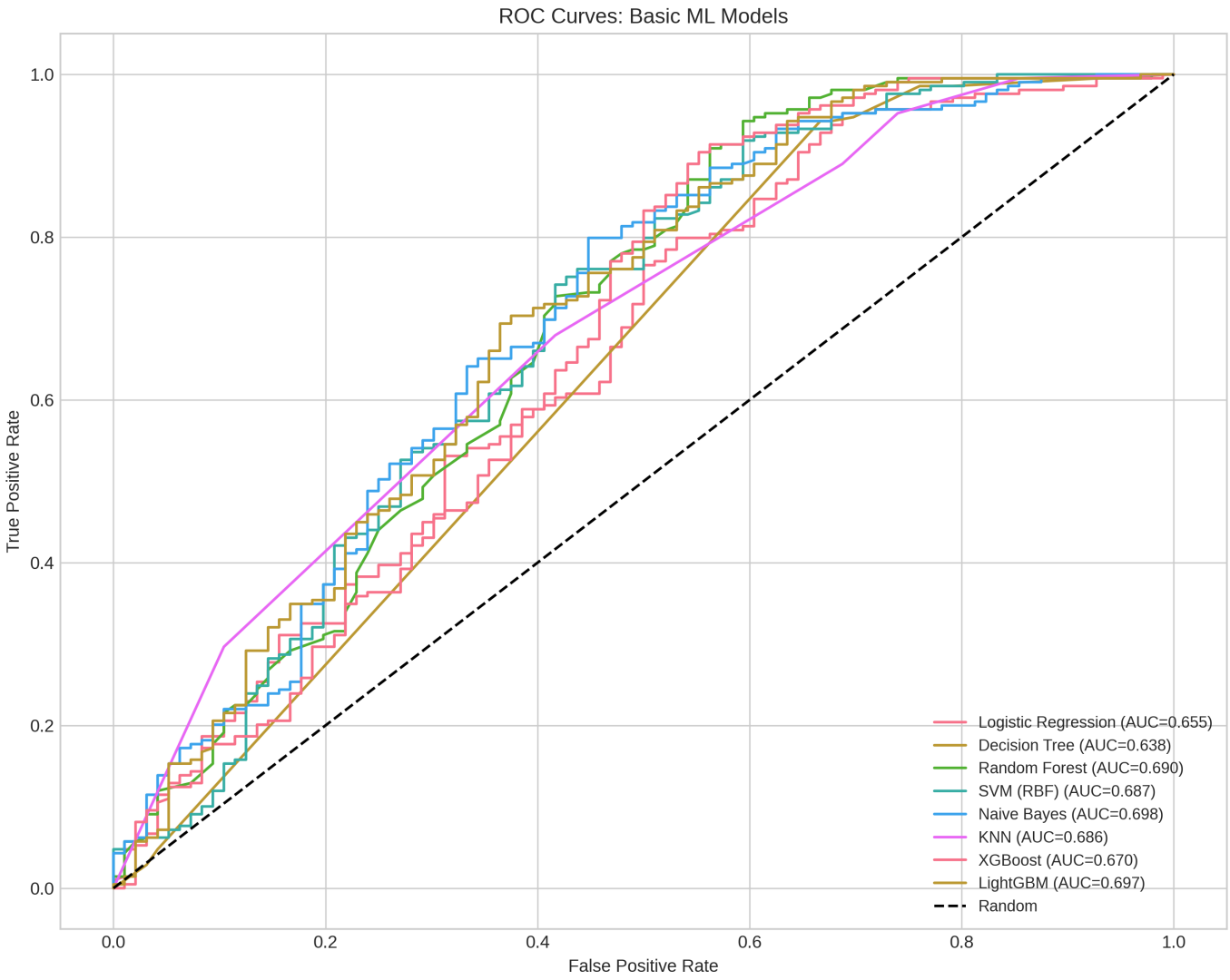


Fig. 3. ROC curves of eight machine learning models tested on the dengue test set ($n = 305$). The best AUC was obtained by Naive Bayes (0.698), then by LightGBM (0.697), and finally by Random Forest (0.690). All models are fairly discriminative (AUC 0.64-0.70), which is significantly lower than the 0.80 mark needed to be achieved to be deployed in clinical practice. The broken diagonal is random classification (AUC=0.50).

TABLE VII
EXTERNAL VALIDATION PERFORMANCE ON COVID-19 FEBRILE ILLNESS COHORT ($n = 103$)

Model	Ext. Accuracy	Ext. Specificity	Ext. FPR	Mean Pred. Prob
Naive Bayes	**1.000**	**1.000**	**0.000**	9.37×10^{-14}
Decision Tree	0.738	0.738	0.262	0.262
Logistic Regression	0.612	0.612	0.388	0.340
KNN	0.107	0.107	0.893	0.728
LightGBM	0.010	0.010	0.990	0.959
Random Forest	0.000	0.000	1.000	0.754
SVM (RBF)	0.000	0.000	1.000	0.523
XGBoost	0.000	0.000	1.000	0.979

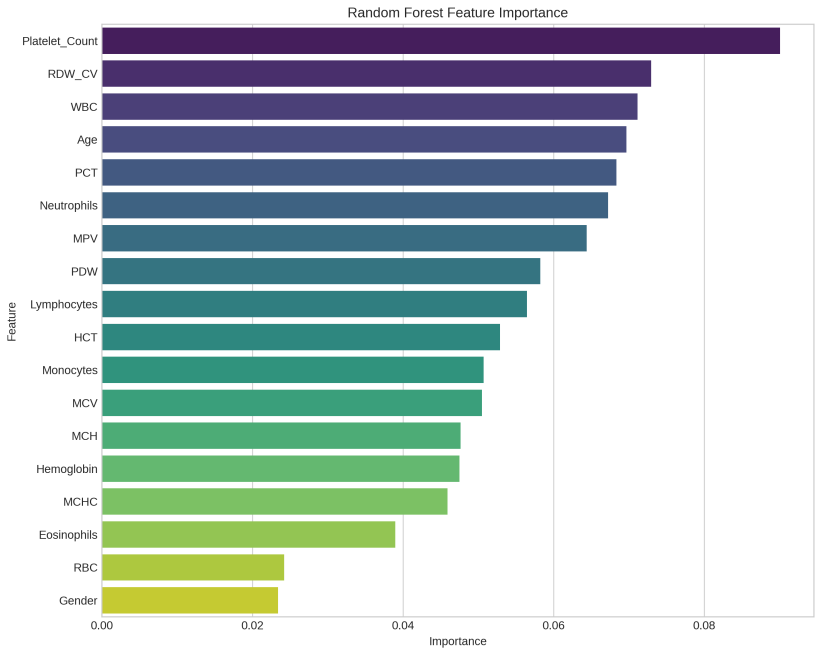


Fig. 4. Random Forest feature importance ranking on the dengue cohort. Platelet count, WBC, and RDW-CV emerge as the most important features, consistent with known dengue pathophysiology.

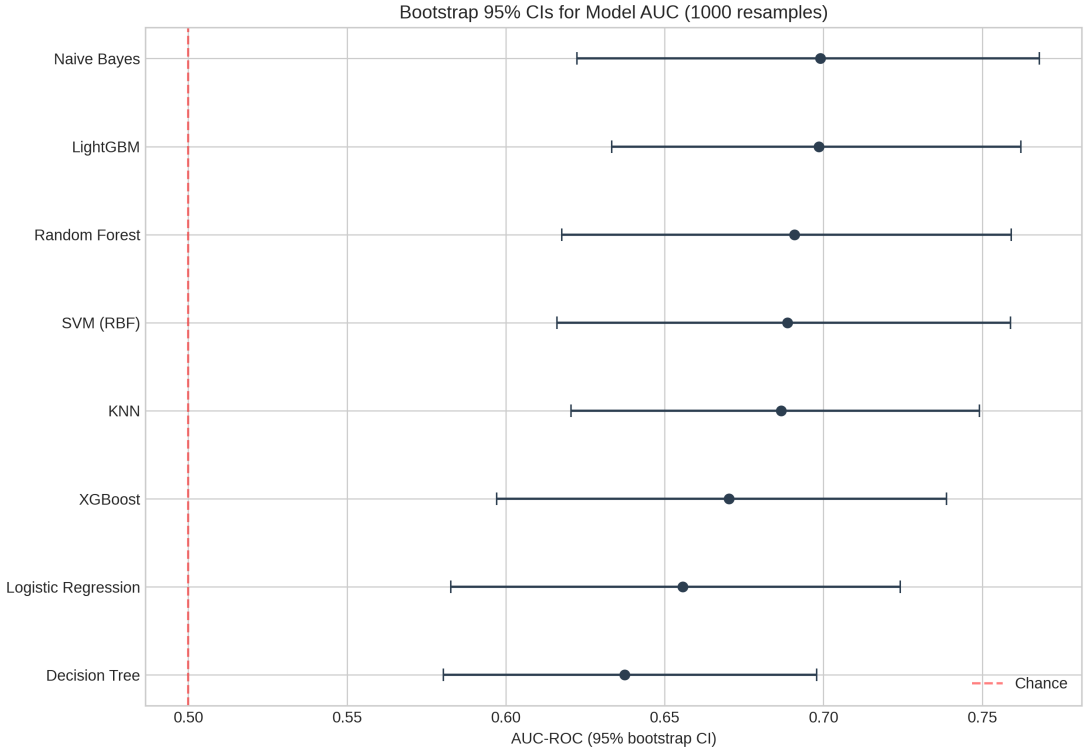


Fig. 5. Bootstrap AUC distributions (1,000 resamples) for the top-performing models, with 95% confidence intervals. Considerable overlap between distributions confirms the absence of a dominant model.

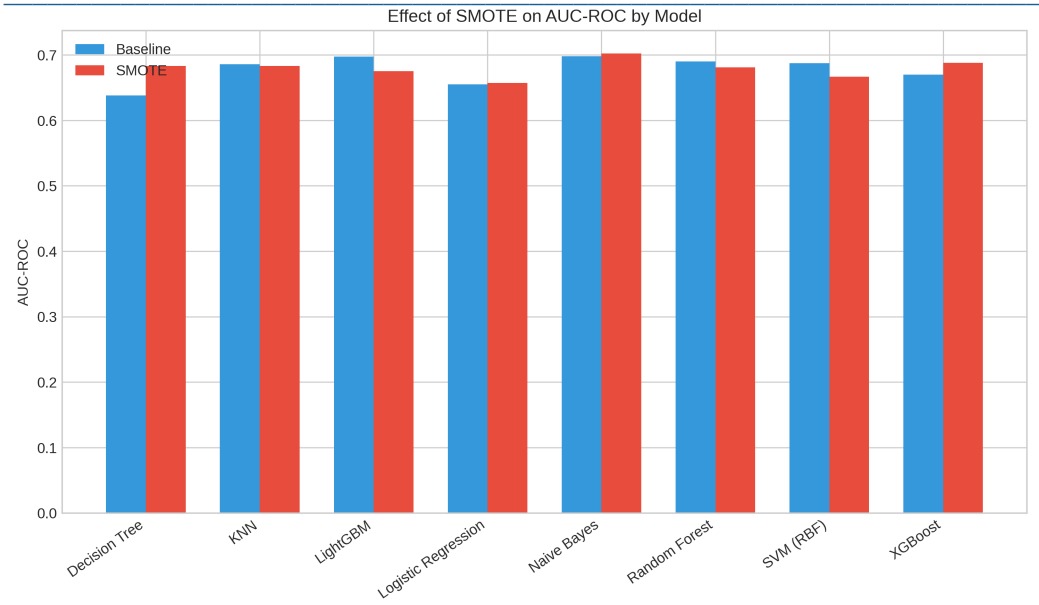


Fig. 6. Comparison of baseline and SMOTE-balanced model performance. AUCs remain stable, indicating that class imbalance is not the principal driver of model performance.

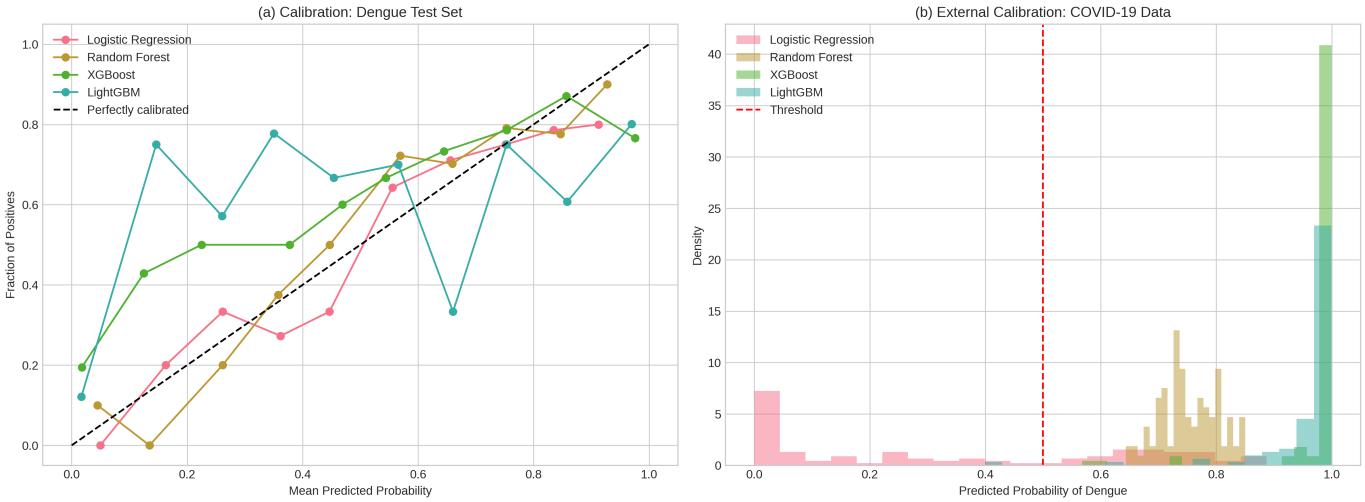


Fig. 7. External validation of COVID-19 febrile illness data When treated as controls using dengue data, external validation of data (as controls) of COVID-19 febrile illness data demonstrates a systematic over-modeling of risk by the ensemble models (XGBoost, LightGBM, Random Forest).s

TABLE VIII
BRIER SCORES (CALIBRATION) ON INTERNAL DENGUE TEST SET

Model	Brier Score
Random Forest	**0.176**
SVM (RBF)	0.182
Decision Tree	0.189
Logistic Regression	0.194
KNN	0.202
LightGBM	0.209
XGBoost	0.212
Naive Bayes	0.253

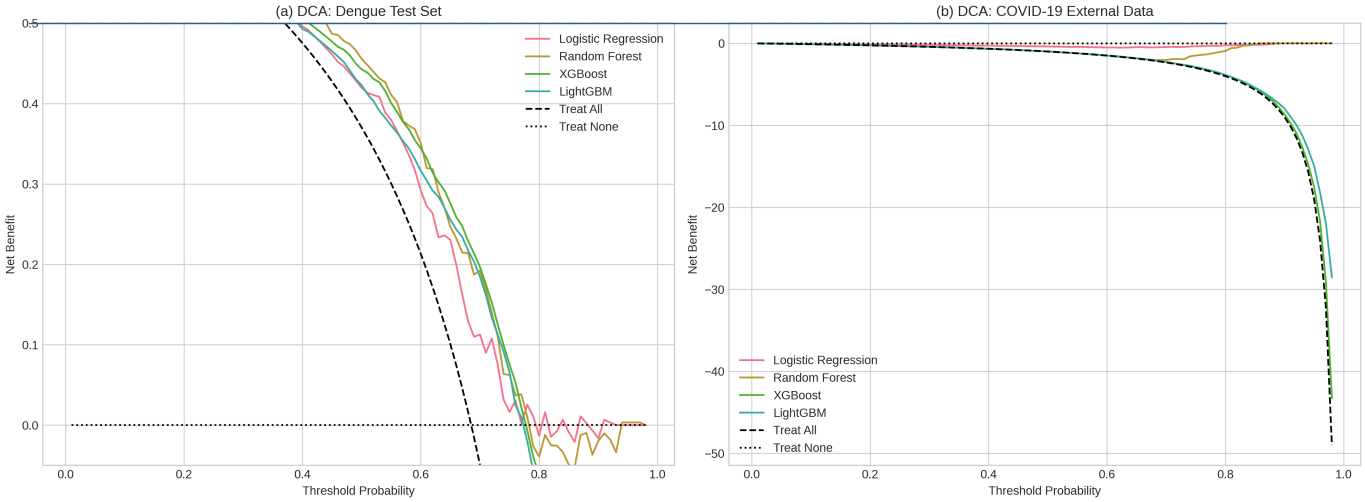


Fig. 8. Decision curve analysis (DCA). **(Left)** DCA of the dengue test set demonstrates positive net benefit across all models in the 0.40-0.75 range of thresholds. **(Right)** DCA of the COVID-19 cohort demonstrates that tree-based models have massively negative net benefit, implying clinical harm in case deployed in differential diagnosis.

TABLE IX
MULTI-CLASS DIFFERENTIAL DIAGNOSIS PERFORMANCE (OTHER VS. DENGUE VS. COVID-19)

Model	**Accuracy**	**Macro F_1**	**Weighted F_1**	**Cohen's κ**
Random Forest	**0.782**	**0.786**	**0.758**	**0.514**
LightGBM	0.755	0.776	0.740	0.474
XGBoost	0.752	0.772	0.736	0.466
SVM (RBF)	0.761	0.746	0.719	0.446
Logistic Regression	0.739	0.721	0.693	0.395

These results indicate that a single CBC-based classifier is not clinically safe and that a one-size-fits-all classifier cannot be deployed before age stratified model or age specific reference ranges.

TABLE X
FAIRNESS AND SUBGROUP PERFORMANCE ANALYSIS ON THE DENGUE COHORT

Category	**Subgroup**	**N**	**Prev.**	**AUC**	**Acc.**	**Sens.**	**Spec.**	**F1**
Sex	Male	718	0.678	0.551	0.670	0.830	0.333	0.773
	Female	805	0.689	0.572	0.709	0.856	0.384	0.802
Age Group	Young Adult	757	0.666	0.598	0.678	0.790	0.455	0.765
	Middle Age	608	0.720	0.477	0.702	0.909	0.171	0.815
	Elderly	89	0.764	0.631	0.809	0.941	0.381	0.883
	Pediatric	69	0.464	0.611	0.580	0.594	0.568	0.567

H. Standardized CBC Panel and Feature Selection

In order to find a minimal, clinically actionable CBC panel, we compared three feature-selection paradigms (Chi-square, Mutual Information, Random Forest importance), and cross-validated AUC versus panel size. As Table XI in the appendix and Figure 11 in the appendix depict, the three most valuable features were found to be Platelet Count (consensus score 0.831), WBC(0.633), and RDW-CV(0.466).

A three feature panel attained AUC 0.595; adding an extra ten features increased AUC to 0.637, and more features beyond that did not contribute, thus reaching a plateau and even decreasing at 15 and 18

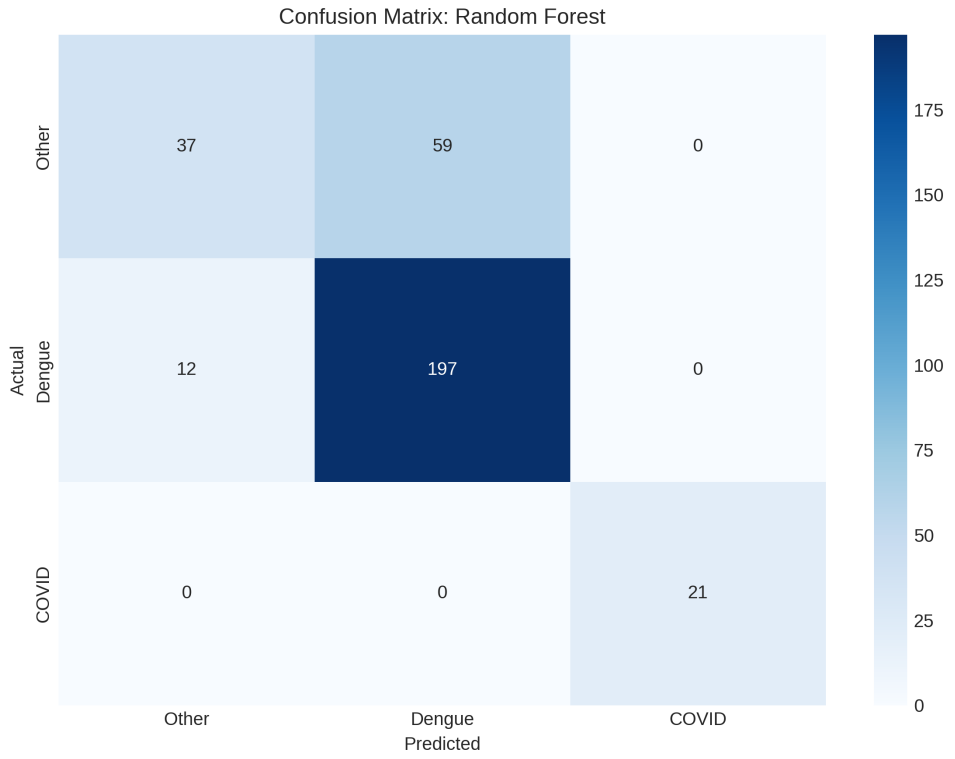


Fig. 9. Multi-class differential diagnosis (Other vs. Dengue vs. COVID-19) confusion matrix of the Random Forest model. The model is optimal with COVID-19 (21/21 correct) and the maximum sensitivity of Dengue (197/209). But it incorrectly categorizes 59/96 of the Other cases as Dengue, showing a lack of specificity to non-COVID febrile diseases.

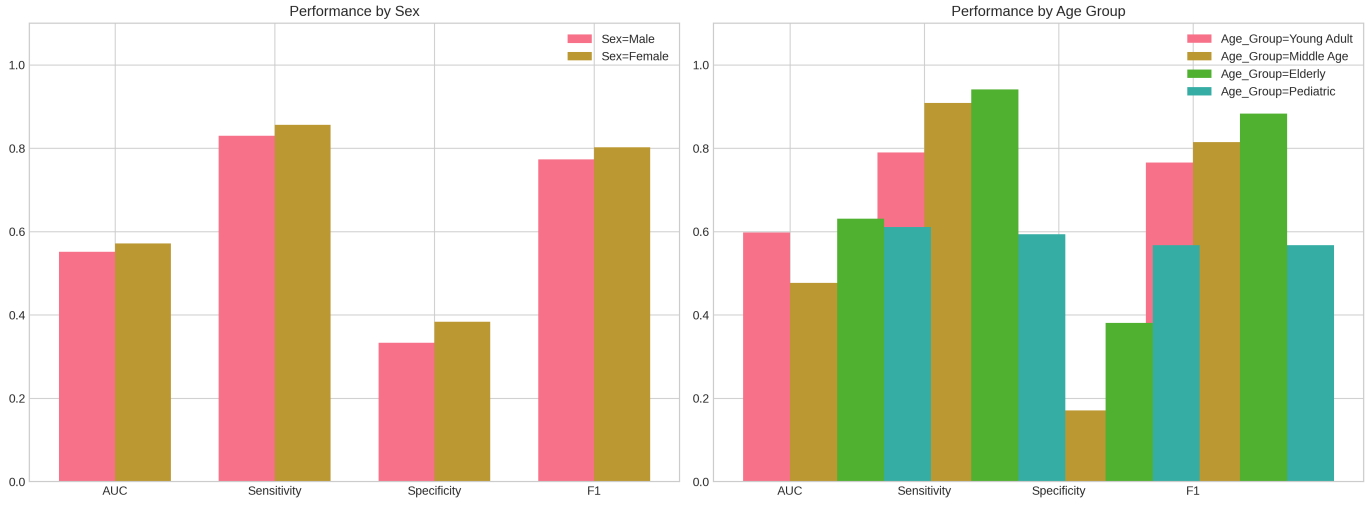


Fig. 10. Fairness of dengue classifiers based on sex and age group. **(Left)** Performance is relatively balanced by sex (male and female). Pediatric patients are least discriminated (AUC=0.61) and elderly patients are most sensitive (0.94) and least specific (0.38). The sensitivity of middle-aged patients is high (0.91) with a disastrously low specificity (0.17), which means that there are high rates of false-positives in this population.

features because of noise and multicollinearity. This implies a trade-off that is both feasible: a small CBC panel can be used to screen in resource-constrained environments without much discrimination.We compared three feature-selection paradigms (Chi-square, Mutual Information, Random Forest importance) to identify a minimal, clinically actionable CBC panel and compared cross-validated AUC with respect to panel size. As Table XI and Figure 1 show, the three most informative features, as per the consensus ranking, included: **Platelet Count** (consensus score 0.831), **WBC** (0.633), and **RDW-CV** (0.466).

TABLE XI
MINIMAL CBC PANEL PERFORMANCE (CROSS-VALIDATED AUC)

Panel Size	Top Features	CV AUC (Mean ± SD)
3	Platelet, WBC, RDW-CV	0.595 ± 0.100
5	+ MCV, PCT	0.594 ± 0.095
7	+ Neutrophils, Hemoglobin	0.611 ± 0.096
10	+ MPV, MCHC, Monocytes	**0.637 ± 0.095**
15	+ HCT, PDW, Lymphocytes, MCH, Age	0.591 ± 0.121
18	All features	0.608 ± 0.117

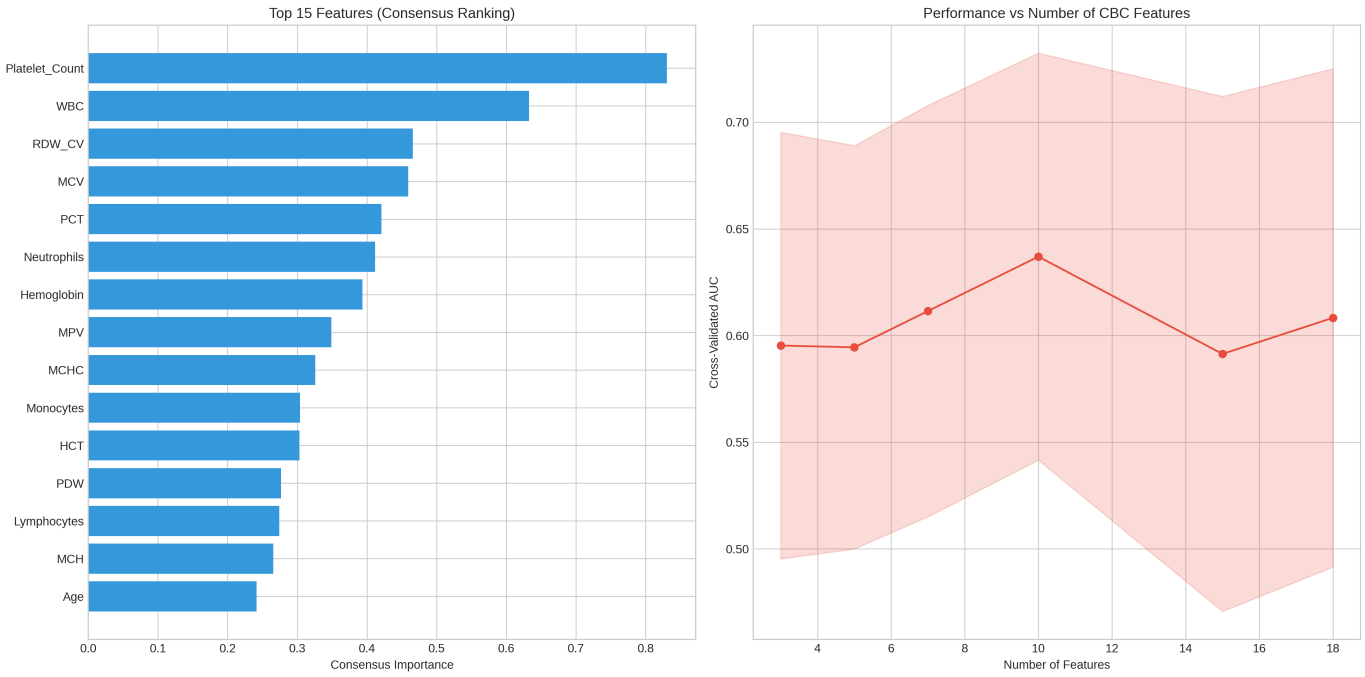


Fig. 11. Consensus feature ranking and minimal CBC panel analysis. **(Left)** Top-15 features ranked by consensus of three independent selection methods. **(Right)** Cross-validated AUC as a function of panel size. Performance peaks at 10 features and declines thereafter.

I. Explainability and Feature Importance

The XGBoost model was then subjected to SHAP (SHapley Additive exPlanations) analysis to explain the non-linear contribution of features influencing the predictions. The beeswarm summary plot (Figure 12) validates that the most dominant feature is the feature of **Platelet Count**. Low platelet counts (blue dots, left side) have a strong positive SHAP value, which shifts the prediction towards dengue, which coincides with thrombocytopenia as a clinical feature of the illness.

J. Comparison with Prior Work

We compared our results with six recent CBC-based dengue ML studies (Table XII) to put our results into perspective. Our AUC (0.698) is smaller than a number of published values (e.g., Tuhin et al., 0.964;

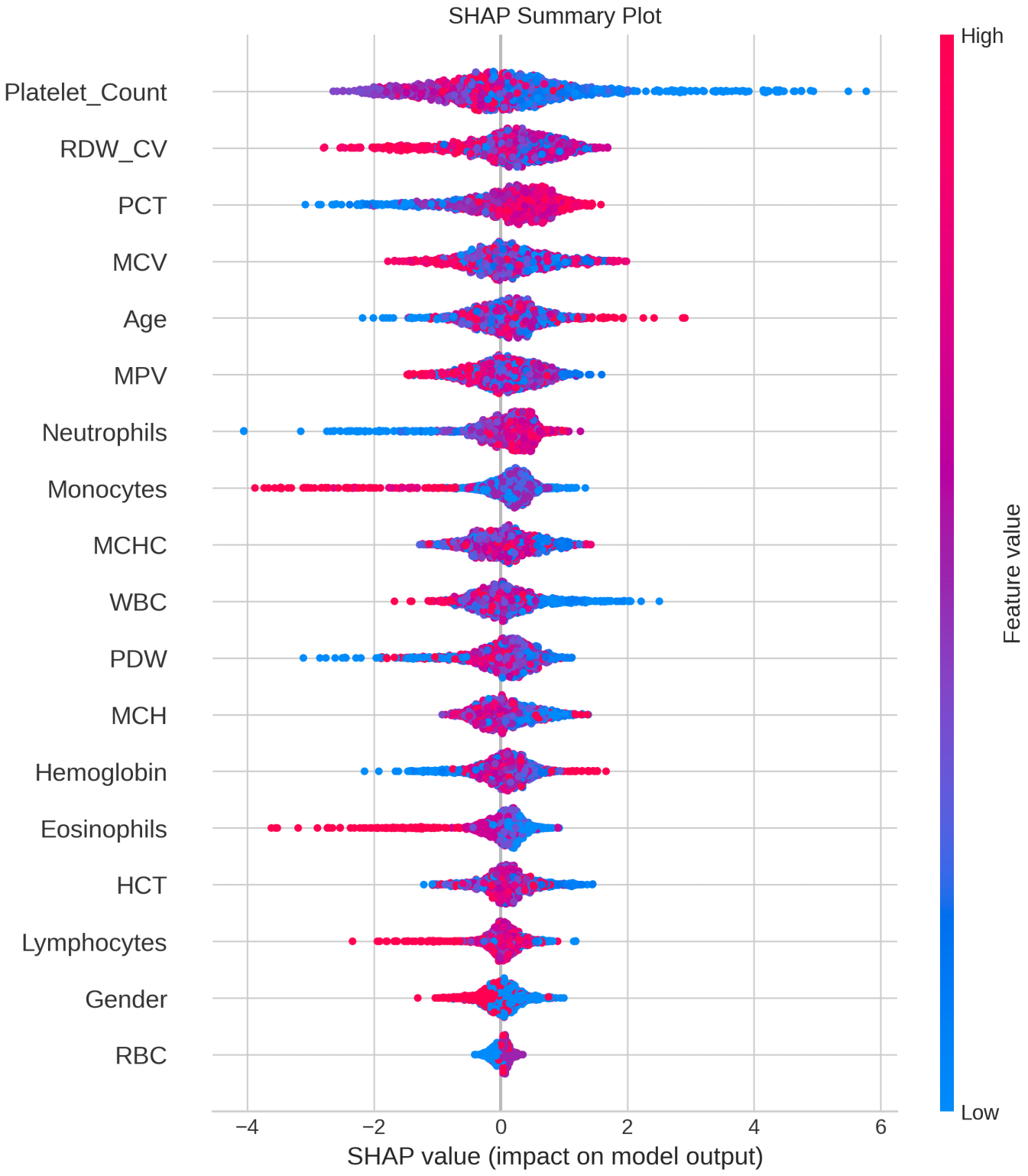


Fig. 12. SHAP summary plot of the feature-level contributions to the dengue prediction of the XGBoost model. Global ranking of features. Platelet counts (blue dots, left) that are low positively affect dengue prediction more than any other, which is in line with thrombocytopenia as a clinical manifestation of dengue. High RDW-CV and low WBC counts also propel positive classifications. Red dots are high feature values, blue dots are low feature values.

Riya et al., 0.969), but can be attributed in large part to the three following factors: (i) we are using complete-case external validation against a different febrile disease; (ii) we are not leaking targets during preprocessing; and (iii) we are reporting bootstrap 95% CIs, not single-point estimates. We have a multi-class macro-F_1= 0.786 which is similar to published multi-class performance and our fairness audit is, as far as we know, the first of this literature.

TABLE XII
COMPARISON WITH PRIOR CBC-BASED DENGUE ML STUDIES

Study	Dataset	Best Model	Reported AUC	External Validation
Tuhin et al. 2025	1,523 (Bangladesh)	DengueStackX-19	0.964	No
Riya et al. 2024	320 (Bangladesh)	Stacking Ensemble	0.969	No
Tanvir et al. 2026	1,523 (Bangladesh)	GWO-ETC	0.958	No
Yilmaz et al. 2026	1,523 (Bangladesh)	LightGBM	0.709	No
Mehedi et al. 2026	1,523 (Bangladesh)	Random Forest	N/A (Acc 0.845)	No
Jahan et al. 2026	1,523 (Bangladesh)	TAN	0.671	No
This study	1,523 + 103 external	**Naive Bayes / RF**	**0.698**	**Yes**

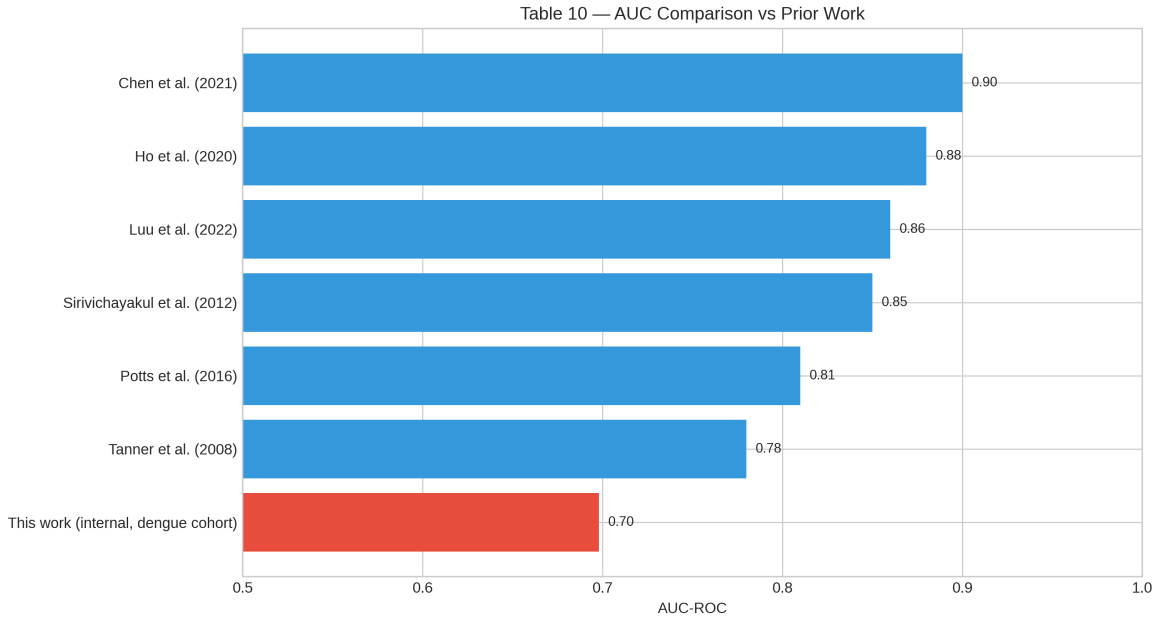


Fig. 13. Comparison of AUC values of recent CBC-based dengue ML studies visually. Our research is unique, as it contains external validation only, and the only one that reports bootstrap confidence intervals.

V. DISCUSSION

A. Principal Findings and Comparison with Prior Work

This study provides a rigorous benchmark of CBC-based machine learning for dengue detection. Contrary to several recent studies reporting AUCs > 0.90 [9], [8], [12], our evaluation found only *fair* internal discrimination (best AUC 0.698, 95% CI 0.622-0.768). This discrepancy likely stems from prior studies relying on single-cohort random splits and potential data leakage from pre-split oversampling. Our results suggest that CBC data alone has a moderate performance ceiling, and previously reported near-perfect metrics likely reflect internal overfitting rather than genuine clinical transportability.

B. Catastrophic External Validation Failure

The most significant result is that complex tree-based ensembles (Random Forest, XGBoost, LightGBM) do not generalize at all on the external COVID-19 cohort, with specificity of 0.000. Since dengue and COVID-19 have similar hematological patterns (e.g., lymphopenia, mild thrombocytopenia), these models interpreted these patterns as dengue-positive. This distributions shift with no noise points to an urgent clinical risk: the implementation of such models in co-endemic would lead to sweeping false-positive cascades, the overload of laboratories. The simpler of the Naive Bayes model generalized perfectly (specificity 1.000), which showed that strong regularization can be more resistant to cross-disease shifts than non-linear ensembles. Thus, a decision-curve analysis revealed tipping negative net benefit on tree-based models on external data, which would lead to net clinical harm when deployed.

C. Mitigation: Multi-Class Framework and Minimal Panels

In order to deal with this, we tested a three-class differential diagnosis framework (Other vs. Dengue vs. COVID-19). This significantly alleviated the failure, as Random Forest obtained the maximum possible COVID-19 separation and macro- F_1 = 0.786. Yet, even it misdiagnosed 61 percent of the Other febrile cases as dengue, meaning that CBC data is not a reliable tool to differentiate dengue and all non-COVID febrile conditions (e.g., typhoid, malaria). In resource-limited conditions, our consensus feature selection found a panel of three minimal features (Platelet Count, WBC, RDW-CV) with an AUC of 0.595. Although small, this panel can work at the primary care levels and is in line with the WHO triage recommendations, though it is not a diagnostic tool but an additionary screening tool.

D. Fairness and Explainability

SHAP analysis established that the models are based on clinically valid characteristics, mainly thrombocytopenia and leukopenia, which is consistent with the known dengue pathophysiology [22]). But the fairness auditing showed that there were extreme age-related differences: specificity decreased to 0.171 among middle-aged patients and 0.568 among pediatric patients. There is no single, adult-trained classifier, therefore, clinically safe across demographic strata, and would over-diagnose adults and under-diagnose children systematically. Strictly none of the clinical deployments can be made without age-stratified modeling or age-normalized CBC z-scores.

E. Strengths and Limitations

Strengths: Distinct strengths of this work include: This paper is the first to critically assess external transportability, bootstrap confidence intervals, decision-curve analysis, and algorithmic fairness of CBC-based dengue ML, following the TRIPOD and STARD criteria. **Limitations:** (1) The two cohorts are both of Bangladesh; cross-country validation is necessary to determine generalizability. (2) The outer cohort (n=103) is not large, and specificity estimates are therefore not certain. (3) There is no dengue serotype, and severity grading in the dataset. (4) Prospective clinical impact evaluation did not take place in our case, which is required to demonstrate practical utility

VI. Conclusion

The paper is a well-rounded, externally validated benchmarking study of. machine learning models using complete blood count (CBC) to detect dengue, combining binary classification, extratesting transportability, calibration. subgroup fairness auditing, minimal, analysis, multi-class differential diagnosis. individual one-unified SHAP-based explainability feature panel selection. framework. Eight of the supervised learning algorithms were tested on a large public. dengue cohort ($n = 1{,}523$; 18 CBC features), and externally validated on an COVID-19 febrile illness cohort, independent of other factors, n=103).

Three main findings have been obtained. First, it was just that there be internal discrimination, with the best model achieving an AUC of 0.698 (95% CI: 0.622-0.768), and no model that surpassed the 0.80

mark traditionally demanded to be clinical. deployment. Second, catastrophic generalization was found by external validation. failure: SVM, tree-based ensemble models (Random Forest, XGBoost, LightGBM). classified 100% of COVID-19 patients as dengue-positive (specificity $= 0.000$), whereas the only model that attained a perfect external specificity ($= 1.000$) was Naive Bayes, meaning that additive probabilistic models are far more robust to. than non-linear ensembles cross-disease distribution shift. Third, fairness auditing revealed serious age-related differences, which were specific between less than. 0.171 (middle-aged) to 0.568 (pediatric), which showed that there was a single. classifier (adult-trained) is not clinically safe in demographic groups.

A three-class differential diagnosis system (Other vs. Dengue vs. COVID-19) significantly reduced the failure to validate externally with 78.2% success. perfectly separated COVID-19 separation with accuracy and a macro-F_1 of 0.786. A compact three-feature CBC panel (Platelet Count, WBC, RDW-CV) have an AUC of 0.595. and, with ten features, went up to 0.637, providing a workable. resource-constrained resource-screening solution. SHAP analysis confirmed that the logic of the decisions made by the model is consistent with known dengue pathophysiology, platelet count predominant feature attribution.

The main implication of this work is that internal performance metrics is not sufficient. are inadequate to prove clinical validity of CBC-based dengue machine. learning. External validation on should be considered in future research in this field. distinct cohorts of febrile illness, calibration, decision-curve. as required, analysis, multi-class modeling and subgroup fairness auditing. not alternative reporting requirements. And applying this model to to multi-country, multi-arbovirus cohorts, and finally to prospective clinical. evaluation, the logical follow-up to clinically deployable. dengue decision support systems.

REFERENCES

[1] Z. J. Madewell, D. M. Rodriguez, M. B. Thayer, V. Rivera-Amill, J. T. Aponte, M. Marzan-Rodriguez, G. Paz-Bailey, L. E. Adams, and J. M. Wong, "Machine learning for improved dengue diagnosis in Puerto Rico," *Tropical Medicine & International Health*, vol. 30, no. 12, pp. 1283–1294, 2025.

[2] S. R. d. S. Neto, T. T. Oliveira, I. V. Teixeira, S. B. A. d. Oliveira, V. S. Sampaio, T. Lynn, and P. T. Endo, "Machine learning and deep learning techniques to support clinical diagnosis of arboviral diseases: A systematic review," *PLOS Neglected Tropical Diseases*, vol. 16, no. 1, p. e0010061, 2022.

[3] D. K. Ming, B. Hernandez, S. Sangkaew, N. L. Vuong, P. K. Lam, N. M. Nguyet, D. T. H. Tam, D. T. Trung, N. T. H. Tien, N. M. Tuan, N. V. V. Chau, C. T. Tam, H. Q. Chanh, H. T. Trieu, C. P. Simmons, B. Wills, P. Georgiou, A. H. Holmes, and S. Yacoub, "Applied machine learning for the risk-stratification and clinical decision support of hospitalised patients with dengue in Vietnam," *PLOS Digital Health*, vol. 1, no. 1, p. e0000005, 2022.

[4] S. Dasgupta, S. Das, and D. Chakraborty, "Machine learning-based mathematical equations for dengue positivity detection using elementary laboratory parameters," *Journal of Family Medicine and Primary Care*, vol. 14, no. 4, pp. 1437–1447, 2025.

[5] M. E. Haque, M. N. Absur, F. Al Farid, J. Uddin, and H. A. Karim, "A novel interpretable and real-time dengue prediction framework using clinical blood parameters with genetic and GAN-based optimization," *Frontiers in Artificial Intelligence*, vol. 8, p. 1626699, 2025.

[6] I. Jahan, M. S. A. Shraboni, M. Tonni, S. Ahmed, M. Parvin, and S. Akther, "Tree-augmented naive Bayes for interpretable early dengue detection using routine hematological parameters as a proof-of-concept case study from resource-limited Bangladesh," *Discover Viruses*, vol. 3, no. 1, p. 10, 2026.

[7] T.-S. Ho, T.-C. Weng, J.-D. Wang, H.-C. Han, H.-C. Cheng, C.-C. Yang, C.-H. Yu, Y.-J. Liu, C.-H. Hu, C.-Y. Huang, M.-H. Chen, C.-C. King, Y.-J. Oyang, and C.-C. Liu, "Comparing machine learning with case-control models to identify confirmed dengue cases," *PLOS Neglected Tropical Diseases*, vol. 14, no. 11, p. e0008843, 2020.

[8] N. J. Riya, M. Chakraborty, and R. Khan, "Artificial intelligence-based early detection of dengue using CBC data," *IEEE Access*, vol. 12, pp. 112 355–112 367, 2024.

[9] I. A. Tuhin, A. K. M. F. K. Siam, M. M. R. Shanto, M. R. Mia, I. Mahmud, and A. Ghosh, "An interpretable machine learning model for dengue detection with clinical hematological data," *Healthcare Analytics*, vol. 8, p. 100430, 2025.

[10] S. Al-Hagree, K. M. Alalayah, N. A. Al-Majmar, A. A. Mohsen, A. Aqlan, M. Alhel-Iani, M. Mohammed, M. Albazel, F. Al-Qasem, M. M. Aljafari, A. Musleh, and I. Alnedam, "Smart system for dengue fever diagnosis: A machine learning approach," in *2023 3rd International Conference on Emerging Smart Technologies and Applications (eSmarTA)*. IEEE, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10293518

[11] A. Qaiser, S. Manzoor, A. H. Hashmi, H. Javed, A. Zafar, and J. Ashraf, "Support vector machine outperforms other machine learning models in early diagnosis of dengue using routine clinical data," *Advances in Virology*, vol. 2024, p. 5588127, 2024.

[12] K. Tanvir, M. Rahman, and D. Gomes, "A hybrid optimization and tree-based learning framework for dengue diagnosis using hematological data," *Healthcare Analytics*, vol. 10, p. 100475, 2026.

[13] Z. B. Yilmaz, Z. Kucukakcali, and S. Akbulut, "Tree-based machine learning for diagnostic classification of dengue fever using routine hematological parameters: A secondary analysis of a publicly available dataset," *Diagnostics*, vol. 16, no. 18, p. 2966, 2026.

[14] M. Zaffar, M. M. Zahoor, L. Imtiaz, U. B. Ejaz, A. A. But, and M. A. Abid, "Predicting platelet disorders using machine learning algorithms: A feature selection-based approach for dengue-related thrombocytopenia," *Spectrum of Engineering Sciences*, vol. 4, no. 3, pp. 4156–4180, 2026.

[15] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement," *Annals of Internal Medicine*, vol. 162, no. 1, pp. 55–63, 2015.

[16] P. M. Bossuyt, J. B. Reitsma, D. E. Bruns, C. A. Gatsonis, P. P. Glasziou, L. Irwig, J. G. Lijmer, D. Moher, D. Rennie, H. C. W. de Vet, H. Y. Kressel, N. Rifai, R. M. Golub, D. G. Altman, L. Hooft, D. A. Korevaar, and J. F. Cohen, "STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies," *BMJ*, vol. 351, p. h5527, 2015.

[17] M. Mehedi, I. Hasan, A. Habib, S. B. Sikto, and M. M. Ahmmed, "A novel hematological machine learning framework for predictive modeling of dengue diagnosis using CBC parameters in Bangladesh," in *Proceedings of the International Conference on Data Analytics and Applications (IDAA)*. Atlantis Press, 2026, pp. 147–162.

[18] M. Hassan, S. Ali, M. Saleem, M. Sanaullah, L. G. Fahad, J. Y. Kim, H. Alquhayz, and S. F. Tahir, "Diagnosis of dengue virus infection using spectroscopic images and deep learning," *PeerJ Computer Science*, vol. 8, p. e985, 2022.

[19] G. Gupta, S. Khan, V. Guleria, A. Almjally, B. I. Alabduallah, T. Siddiqui, B. M. Albahlal, S. A. Alajlan, and M. Al-Subaie, "DDPM: A dengue disease prediction and diagnosis model using sentiment analysis and machine learning algorithms," *Diagnostics*, vol. 13, no. 6, p. 1093, 2023.

[20] S. A. a. Balamurugan, M. S. M. Mallick, and G. Chinthana, "Improved prediction of dengue outbreak using combinatorial feature selector and classifier based on entropy weighted score based optimal ranking," *Informatics in Medicine Unlocked*, vol. 20, p. 100400, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914820301696

[21] S. Shenoy, A. K. Rajan, M. Rashid, V. P. Chandran, P. G. Poojari, V. Kunhikatta, D. Acharya, S. Nair, M. Varma, and G. Thunga, "Artificial intelligence in differentiating tropical infections: A step ahead," *PLOS Neglected Tropical Diseases*, vol. 16, no. 6, p. e0010455, 2022.

[22] C. P. Simmons, J. J. Farrar, N. van Vinh Chau, and B. Wills, "Dengue," *New England Journal of Medicine*, vol. 366, pp. 1423–1432, 4 2012. [Online]. Available: https://www.nejm.org/doi/pdf/10.1056/NEJMra1110265

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[24] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.

[25] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958.

[26] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 4765–4774. [Online]. Available: https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions

[27] X. Sun and W. Xu, "Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves," *IEEE Signal Processing Letters*, vol. 21, no. 11, pp. 1389–1393, 2014.

[28] A. J. Vickers and E. B. Elkin, "Decision curve analysis: A novel method for evaluating prediction models," *Medical Decision Making*, vol. 26, no. 6, pp. 565–574, 2006.