A Novel Hematological Machine Learning Framework for Predictive Modeling of Dengue Diagnosis Using CBC Parameters in Bangladesh

Md Mehedi Imam Hasan[1*], Ahasan Habib[1], Sawhardo Biswas Sikto[1], and Md. Mortuza Ahmmed[2]

[1] Department of Computer Science and Engineering,
American International University-Bangladesh (AIUB), Dhaka-1229, Bangladesh
[2] Department of Mathematics,
American International University-Bangladesh (AIUB), Dhaka-1229, Bangladesh
22-47413-2@student.aiub.edu *,
22-48877-3@student.aiub.edu,
22-47406-2@student.aiub.edu,
mortuza@aiub.edu

Abstract. Early diagnosis of dengue fever is critical to patient care and outbreak control, yet reliable clinical indicators are often elusive. Machine learning (ML) offers promise by learning complex patterns in clinical and hematological data. In this study, we first review recent ML-based dengue diagnosis models that use routine clinical or blood parameters. Notably, Support Vector Machines (SVM) and Random Forests (RF) frequently perform best. Key studies include a Brazilian RF model (85% accuracy) for misdiagnosed hospital cases and a Brazilian symptom-based screening model (93% accuracy) using decision trees and neural networks. However, most work focuses on non-Bangladeshi cohorts, with few models addressing explainability. We then apply ML to a new dataset of 1,523 Bangladeshi patients with complete blood count (CBC) and demographic features. We preprocess and encode features (e.g., one-hot gender) and split data into training/test sets. We train logistic regression, RF, and XGBoost models and evaluate accuracy, F1-score, and ROC AUC. Random forest performed best (84.5% accuracy, AUC 69%), outperforming logistic regression (61.2% accuracy), XGBoost (81% accuracy), SVM (81.4%) and neural network (81%). These results highlight the potential of ML for dengue screening in underrepresented populations. The findings underscore gaps in local data, limited feature sets, and the need for interpretable models (e.g., using SHAP/LIME) to support clinicians.

Keywords: Dengue fever, hematological parameters, machine learning, predictive model, Random Forest, Support Vector Machines, Bangladesh.

1 Introduction

Dengue stands as one of the most common mosquito-borne diseases on the planet since it leads to roughly one hundred million to four hundred million

M. S. Arefin et al. (eds.), *Proceedings of the International Conference on Intelligent Data Analysis and Applications (IDAA 2025)*, Advances in Intelligent Systems Research 206,
https://doi.org/10.2991/978-94-6239-664-7_12

infections annually. [1]. The World Health Organization identified dengue as the second most significant infectious disease worldwide, behind COVID-19, during its 2021 assessment, which demonstrated its fast spread and high global impact. spread [2]. The Americas alone experienced more than 7.6 million dengue cases throughout the first four months of 2024, making these outbreaks historically significant. cases [2].Dengue infection presents itself through different degrees of clinical manifestations ranging from simple febrile symptoms to serious cases of dengue hemorrhagic fever that lead to shock and bleeding and organ failure. Specific antiviral treatments are unavailable at present, and severe outcomes commonly result in death when not treated promptly. The detection of dengue must happen at an early stage in order to determine suitable treatment approaches and monitoring procedures. The review of recent literature shows that multiple ML/DL models have been used for dengue prediction, but support vector machines (SVM), random forests (RF), and gradient-boosting models consistently show the best prediction accuracy. [3].These models have achieved accuracy in the 80–95% range on retrospective datasets [4, 5],demonstrating their potential clinical value.

Routine hematological tests can provide early indicators of dengue. Common findings include leukopenia (WBC count $< 5{,}000/\text{mm}^3$), significant thrombocytopenia (platelets $< 150{,}000/\text{mm}^3$), and hemoconcentration (elevated hematocrit) [6]. For example, thrombocytopenia is a hallmark of dengue infection and is used by WHO as a severity indicator [6, 7]. These inexpensive laboratory results are especially valuable in low-resource settings, where rapid point-of-care diagnostics may be unavailable [6,7]. Recent work has shown that machine learning (ML) models can exploit such laboratory features to predict dengue presence and severity. Rao et al. demonstrated that a complete blood count (CBC) can serve as an early prognostic tool for dengue in resource-limited settings [6]. In computational studies, modern ML methods (e.g., neural networks, decision trees) have achieved high accuracy in predicting dengue from clinical data [8]. For instance, Dasgupta et al. used only five laboratory parameters to train ML classifiers, achieving 95.8% accuracy for dengue positivity [7].

Despite these advances, important gaps remain. Few studies have applied ML exclusively to routine hematology data at scale, and class imbalance (more dengue-positive than negative cases) can bias models. Our work addresses these gaps by using a sizeable real-world dataset from Bangladesh, applying SMOTE oversampling to balance classes, and comparing multiple ML algorithms. We emphasize interpretability and clinical utility: all input features are standard CBC parameters, and we evaluate our models according to interpretable metrics. Through this study, we are going to have a better comparison of classifiers and objectively measure their power and limitations, which is in line with previous studiesRao2020,Dasgupta2025.

2 Literature Review

In recent years, multiple studies have utilized machine learning (ML) to identify dengue using patient data. As summarized in Table 1, their approaches and results show common themes. Most studies use classification models (dengue vs. non-dengue) and report metrics such as accuracy, sensitivity/specificity, F1-score, or AUC [3, 4]. Support vector machines (SVM) and tree-based methods (random forest, gradient boosting) are frequently top performers [3, 4], with key features often including platelet count, white blood cell count, hematocrit, and clinical symptoms [5, 9].

For example, Santos et al. (2023) analyzed Brazilian hospitalization data (2014–2020) and built RF and other models to flag cases where dengue might have been misdiagnosed. Their RF model achieved about 85% accuracy in identifying 13,608 hospitalizations potentially miscoded as other diseases [4].Bruhn et al. (2024) used data from Brazil's national registry (2016–2019) including 10,000 confirmed dengue and 10,000 discarded cases with tree-based ML to screen suspected cases. A decision tree and a neural network (multi-layer perceptron) both achieved 92–93% accuracy (AUC 0.99) using ten questionnaire-based features [10].These models showed ML could effectively identify dengue from clinical signs alone. Several studies have explored ML for dengue detection: for example, Bairy et al. used ML on peripheral blood smear features (platelet and lymphocyte morphology) to classify dengue with 95% accuracy [11].Ferreira et al. showed decision trees and neural networks can effectively screen dengue from clinical notification data [10].

In Colombia, Arrubla-Hoyos et al. (2024) developed decision tree and RF classifiers on a dataset of patients with dengue, Zika, or chikungunya (similar arboviral diseases). Using signs, symptoms, and lab results, their RF model classified dengue with 88

Valdez et al. (2024) conducted a case-control study in Mexico using an artificial neural network (ANN) built on operational case definitions. Among 233 confirmed dengue cases and 233 non-dengue controls, their ANN achieved 90% sensitivity and 82% specificity [12],surpassing a rule-based "direct" algorithm. This indicates ML's advantage over standard symptom checklists.

Many studies focus on hematological data. Mayrose et al. (2023) proposed an image-based ML approach: they detected platelet counts and extracted lymphocyte nucleus features from digital blood smears. Using 10 morphological and textural features and various classifiers (SVM, decision tree, etc.), they achieved 93.6% accuracy with simple features and up to 95.7% accuracy using deep features and SVM [13].Although image-based, this work underscores the diagnostic value of automated blood counts (platelets, lymphocytes).

Qaiser et al. (2024) used routine laboratory and serology data from Pakistan to predict RT-PCR–confirmed dengue. Among 300 patients, they trained multiple models (logistic, SVM, XGBoost, LightGBM, RF, and CatBoost). SVM performed best (71.4% accuracy, 97.4% recall) in predicting PCR results [9].This

shows that even with incomplete tests, ML can leverage lab trends for rapid screening.

Overall, these studies achieve high performance (often 80% accuracy) on retrospective data. However, nearly all were conducted in non-Bangladeshi settings. Only a few involved South Asian patients (e.g., Pakistan). As Table 1 shows, only Sarma et al. (Bangladesh) is noted, but with a small sample (n=209) and moderate accuracy (79%) [14].There remains a lack of large studies on Bangladeshi cohorts. Additionally, most models are evaluated only on internal data; external validation is rare. Finally, few studies examine why certain features drive predictions. For instance, the dual-level MLP approach (2025) highlighted the use of SHAP/LIME to interpret predictions [1], but most models remain opaque. Future work should integrate explainable AI to yield clinician-trustworthy decision support.

Table 1: Summary of Key Research Themes

Author	**Technique**	**Result**
Santos et al. (2023)	RF, SVM, LR, DT	RF: 85% accuracy, F1=0.85
Bohm et al. (2024).	DT, MLP, KNN	DT/MLP: 92–93% acc, AUC=0.99
Arrubla-Hoyos et al. (2024)	DT, RF	RF: 88.0% dengue acc; sensitivity 99%, specificity 100%
Mayrose et al. (2023)	SVM, DT, SVM-deep	SVM/DT: 93.6% acc; SVM (100 deep+100 LBP feats): 95.7%
Qaiser et al. (2024)	LR, RF, XGB, SVM, LGBM, CatBoost	SVM: 71.4% acc, 97.4% recall
Bairy et al., (2023)	SVM, RF, neural networks	SVM accuracy 93.62%

Research Gap: Despite these advances, few studies have exploited large CBC-only datasets for dengue screening. Many ML applications focus on imaging or include many clinical factors, whereas our work uses only standard CBC features readily available at primary care. Prior models also often omit explicit handling of class imbalance [15].Our work is motivated by the need for practical, data-driven tools that leverage common blood tests. By highlighting classifier comparisons on a balanced training set and providing detailed metric breakdowns, we aim to clarify which models and features are most effective for dengue prediction. Platelet count, for instance, was previously found to be the most influential single feature [7],a hypothesis we examine in our analysis.

3 Methodology

In the next section, we elaborate on the methodology used to establish a small-volume machine learning guide for rapid dengue diagnosis using commonly avail-

able hematological parameters. The suggested approach, as seen in Fig. 1, is structured as a pipeline that starts with data collection and ethics processes, followed by extensive preprocessing (for example: encoding, imputation, clinical ratios, scaling, and class-imbalance gemming). The next subsection describes how we train multiple supervised classifiers on our feature set, including hyperparameter tuning and cross-validated training. Lastly, we document the model performance metrics and interpretability analyses to highlight clinically interpretable hematological predictors. It offers an equally rigorous, transparent, and reproducible basis for forecasting model tasks.

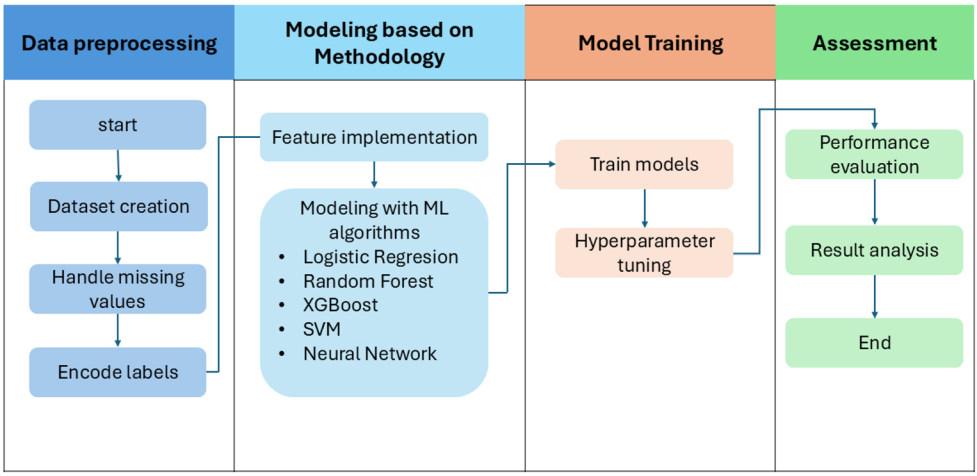

Fig. 1: Workflow of the Data Processing and Model Development Pipeline

3.1 Data Collection

We used the public *"Dengue Fever Hematological Dataset: Clinical Insights for Improved Diagnosis and Patient Management"* (version 2) hosted on Mendeley Data at: https://data.mendeley.com/datasets/6fsrsk3mb8/2.

The dataset contains **1,523 anonymized patient records** collected at *Jamalpur 250-Bedded General Hospital*, Jamalpur, Bangladesh between 10 February 2024 and 27 September 2024. Each row corresponds to a single patient, and includes:

– **Demographic data:** age, gender;

- **Hematological parameters:** hemoglobin, hematocrit, total WBC count, differential leukocyte counts (neutrophils, lymphocytes, monocytes, eosinophils), red blood cell indices, platelet count and related indices;
- **Diagnostic label:** dengue test result (Positive / Negative) determined by standard diagnostic methods, including PCR, NS1 antigen testing, and IgM/IgG serology.

The data are provided in CSV format and are loaded into a pandas DataFrame (`Dengue.csv`). The primary prediction target is the binary variable `Result` (1 = dengue-positive, 0 = dengue-negative). All remaining columns (including engineered ratios, see Section 3.2) are used as candidate predictors for machine learning–based diagnosis and biomarker discovery.

Ethical considerations. The study was performed following the required ethical guidelines and regulations of the institutions. Approval from the hospital institutional review board was obtained, and all records were fully anonymized prior to analysis. The study maintained patient confidentiality and data security.

3.2 Data Preprocessing

The data preprocessing is important for the dataset to allow for robust machine learning modeling. The following operations were performed.

Data cleaning and encoding Scikit-learn's `LabelEncoder` is used to numerically encode categorical variables output:

- `Gender`: 1 for male and 0 for female;
- `Result`: 1 for dengue-positive and 0 for dengue-negative.

As such, they have been structured to be compatible with standard classifiers, such as logistic regression, support vector machines (SVM), tree-based, and neural network-based models.

Handling missing values We didn't have missing values on the main variables in the dataset used in our experiments. But still, to make the results more robust and reproducible, we set up mean imputation `SimpleImputer(strategy="mean")` on all numeric columns. It is simply the imputation fitted on observed data and used to transform the DataFrame to be in an imputed df imputed version. Whereas if there were missing values, these would be substituted with the respective means of the column, preventing the loss of a sample owing to listwise deletion.

Feature engineering To enhance the clinical signal captured by the models, we derive several clinically motivated ratio features based on interactions between hematological parameters known to be important in dengue:

$$\text{Platelet\_HCT\_Ratio} = \frac{\text{Total Platelet Count(/cumm)}}{\text{HCT(\%)}} \tag{1}$$

$$\text{Neutrophil\_Lymphocyte\_Ratio} = \frac{\text{Neutrophils(\%)}}{\text{Lymphocytes(\%)}} \tag{2}$$

$$\text{WBC\_Platelet\_Ratio} = \frac{\text{Total WBC count(/cumm)}}{\text{Total Platelet Count(/cumm)}}. \tag{3}$$

These features capture relationships between inflammation, immune response, hemoconcentration, and thrombocytopenia key aspects of dengue pathology and are appended as new columns to `df_imputed`.

Train–test split and feature scaling For Task 1 (dengue diagnosis prediction), we define:

- Feature matrix X: all columns in `df_imputed` except `Result`;
- Target vector y: encoded `Result` (0 = negative, 1 = positive).

We perform an 80/20 stratified split using `train_test_split(test_size=0.2, stratify=y, random_state=42)` to preserve the original class distribution in both training and test sets and reduce sampling bias.

Many of the employed algorithms (e.g., logistic regression, SVM, neural networks) are sensitive to feature scales. Therefore, we standardize all numeric predictors using scikit-learn's `StandardScaler`:

$$X_{\text{train\_scaled}} = \text{scaler.fit\_transform}(X_{\text{train}}), \qquad X_{\text{test\_scaled}} = \text{scaler.transform}(X_{\text{test}}). \tag{4}$$

All downstream models are trained and evaluated on the standardized features.

Addressing class imbalance with SMOTE The dataset is imbalanced, with approximately 68% dengue-positive and 32% dengue-negative cases. Such imbalance may cause models to favor the majority class.

To mitigate this, we apply Synthetic Minority Over-sampling Technique (SMOTE) to the training set only, using `SMOTE(random_state=42)`. SMOTE generates synthetic instances of the minority class by interpolating between existing minority samples in feature space, resulting in a balanced training distribution:

$$\left(X^*_{\text{train\_scaled}}, y^*_{\text{train}}\right) = \text{SMOTE}\left(X_{\text{train\_scaled}}, y_{\text{train}}\right). \tag{5}$$

The held-out test set remains untouched, ensuring unbiased evaluation.

Exploratory correlation and dimensionality reduction To understand relationships between variables and the structure of the feature space, we perform:

- **Correlation analysis:** A Pearson correlation matrix is computed on `df_imputed` (including engineered features and the target) and visualized with a seaborn heatmap. This highlights strongly associated hematological variables and potential multicollinearity.
- **Principal Component Analysis (PCA):** PCA with two components is applied to the standardized full feature set, yielding a 2D projection of patients. The explained variance ratio of the first two components is reported and visualized, providing insight into dominant directions of variance.
- **t-distributed Stochastic Neighbor Embedding (t-SNE):** t-SNE (2D, perplexity = 30, `random_state=42`) is applied to the same standardized features, and the resulting embeddings are plotted with color-coding by dengue status. This allows visual inspection of potential clusters of dengue-positive vs. dengue-negative patients in a non-linear manifold.

These analyses are exploratory and are not used directly for model training.

3.3 Model Development and Analysis

We formulate the learning task as a supervised binary classification problem to predict dengue diagnosis from hematological profiles (Task 1) and use tree-based models and explainability tools to identify key biomarkers (Task 2).

Model family (Task 1: diagnosis prediction) We implement a diverse set of widely used classifiers to compare linear, ensemble, kernel, and neural approaches:

- Logistic Regression (LR) – linear baseline with L2 regularization;
- Random Forest (RF) – bagged ensemble of decision trees;
- Extreme Gradient Boosting (XGBoost) – gradient-boosted trees;
- Support Vector Machine (SVM) – kernel-based classifier with probabilistic outputs;
- Multilayer Perceptron (MLP) – feed-forward neural network.

All models are trained on the SMOTE-balanced, standardized training data $\left(X^*_{\text{train\_scaled}}, y^*_{\text{train}}\right)$.

Hyperparameter tuning and training setup For each model, we define a hyperparameter grid and perform grid search with 5-fold cross-validation (`GridSearchCV`, `cv=5`, `scoring="accuracy"`, `n_jobs=-1`):

- **Logistic Regression (LR):**

$$C \in \{0.1, 1, 10\} \tag{6}$$

– **Random Forest (RF):**

$$n_{\text{estimators}} \in \{100, 200\}, \qquad \text{max\_depth} \in \{10, 20, \text{None}\} \tag{7}$$

– **XGBoost:**

$$n_{\text{estimators}} \in \{100, 200\}, \quad \text{max\_depth} \in \{3, 5, 7\}, \quad \text{learning\_rate} \in \{0.01, 0.1\}, \tag{8}$$

with `eval_metric = "logloss"`.

– **SVM:**

$$C \in \{0.1, 1, 10\}, \qquad \text{kernel} \in \{\text{linear}, \text{rbf}\}, \tag{9}$$

with probability estimates enabled (`probability = True`) to compute AUC-ROC.

– **Neural Network (MLPClassifier):**

$$\text{hidden\_layer\_sizes} \in \{(100,), (100, 50)\}, \quad \alpha \in \{10^{-4}, 10^{-3}\}, \quad \text{max\_iter} = 1000. \tag{10}$$

The primary optimization criterion during grid search is accuracy. For each model, the best hyperparameter configuration is selected as the one achieving the highest mean cross-validation accuracy.

After grid search:

1. The best estimator is further evaluated with 5-fold cross-validation on the training data to obtain mean and standard deviation of accuracy;
2. The best estimator is retrained on the entire SMOTE-balanced training set and evaluated once on the held-out test set.

3.4 Model Evaluation

Model evaluation is carried out on the held-out test set using the best hyperparameters obtained from the grid search.

Let:

– N: number of test samples;
– $y_i \in \{0, 1\}$: true label for sample i (0 = negative, 1 = positive);
– $\hat{y}_i \in \{0, 1\}$: predicted label for sample i;
– $\hat{p}_i = P(y_i = 1 \mid x_i)$: predicted probability of dengue-positive (when available).

From the confusion matrix, we obtain:

– True Positives (TP) – correctly predicted positives;
– True Negatives (TN) – correctly predicted negatives;
– False Positives (FP) – negatives predicted as positives;
– False Negatives (FN) – positives predicted as negatives.

Accuracy

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = y_i) = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{11}$$

Precision, recall, and F1-score For the positive (dengue-positive) class:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{12}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{13}$$

and

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{14}$$

These metrics are computed with scikit-learn's `precision_score`, `recall_score`, and `f1_score`, using dengue-positive as the positive class.

AUC-ROC Given predicted probabilities $\hat{p}_i$, we construct the Receiver Operating Characteristic (ROC) curve by varying the decision threshold t and computing:

$$\text{TPR}(t) = \frac{\text{TP}(t)}{\text{TP}(t) + \text{FN}(t)}, \tag{15}$$

$$\text{FPR}(t) = \frac{\text{FP}(t)}{\text{FP}(t) + \text{TN}(t)}. \tag{16}$$

The Area Under the ROC Curve (AUC-ROC) is defined as

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}), \tag{17}$$

and is computed numerically via `roc_auc_score(y_test, p̂)`.

Cross-validation accuracy For each model, 5-fold cross-validation yields accuracy scores $\{a_1, a_2, a_3, a_4, a_5\}$. We report:

$$\overline{a} = \frac{1}{5}\sum_{k=1}^{5} a_k, \qquad \sigma_a = \sqrt{\frac{1}{5}\sum_{k=1}^{5}(a_k - \overline{a})^2}. \tag{18}$$

Here, $\overline{a}$ is the mean cross-validation accuracy, and σ_a is the standard deviation, summarizing the stability of each model across folds.

Final reporting We create a comparison table for all five classifiers, ranked from best to worst on the following metrics:

- Test-set Accuracy, Precision, Recall, F1-score, and AUC-ROC;
- Mean ± standard deviation cross-validation accuracy.

In parallel to this table, feature-importance ranks (RF and XG Boost), SHAP-based explanations, and visual diagnostics (correlation matrix, PCA, t-SNE, and accuracy bar plot) are also provided in this table. The top-performing model is represented as the best candidate for interpretable dengue diagnosis and hematological biomarker analysis by a consensus of held-out performance and stability.

3.5 Dataset Description and Limitations

Despite the dataset providing useful real-world information, there are a few limitations we should acknowledge:

1. **Sample size.** The cohort size of 1523 patients is considered moderately small compared with very large-scale studies (e.g., 10,000 patients). It may be limited in statistical power and generalizability.
2. **Class imbalance.** The class imbalance is introduced by the class distribution (about 68% dengue-true vs. 32% dengue-false). This is fixed with SMOTE on the training data, but there may be some remaining bias towards the majority class.
3. **Feature completeness.** It is also limited by clinical data, primarily limited to routine hematological parameters, as detailed clinical symptoms (e.g., fever pattern, rash), imaging data, or more advanced, in some cases, research based biomarkers (e.g., cytokines) are missing from the dataset. Though this allows the approach to be deployable in low-resource settings, it could limit the top performance that can be reached.
4. **Demographic and site bias.** All records come from a single public hospital in Bangladesh. The resulting models may not fully reflect the diversity of other geographic regions, healthcare systems, or demographic groups.
5. **Ethical and practical constraints.** To preserve confidentiality, certain potentially informative variables (e.g., comorbidities, treatment history) were not included. This restricts the scope of the analysis but ensures compliance with ethical requirements.

4 Results

Our model comparison yielded the results summarized presented in Table 2. On the test set after applying k-fold cross validation, random forest achieved the highest balanced performance: about 84.5% accuracy, a 83.8% F1-score for dengue cases, and AUC 69%. In contrast, SVM attained 81.4% accuracy and F1 72%, while XGBoost achieved 80.9% accuracy, F1 82%, and an AUC of

Table 2: Model performance comparison (test set)

Model	Acc. (%)	Recall (%)	F1 (%)	AUC
Random Forest	84.5	91.9	83.8	0.69
XGBoost	80.9	89.5	82.7	0.68
SVM	81.4	80.4	77.6	0.66
Neural Net	80.9	74.2	75.8	0.67
Logistic Reg.	61.3	67.9	72.3	0.54

68%. The logistic regression performed poorly: it achieved only 61.2% accuracy (it predicted almost all cases as dengue due to imbalance). Here is Correlation Matrix of Hematological Features and Result:

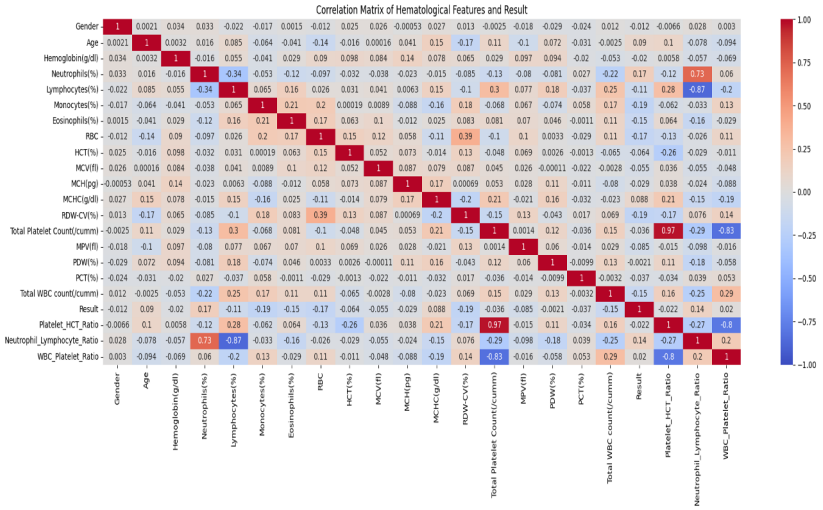

Fig. 2: Correlation matrix of hematological features and dengue result.

A detailed analysis of the correlation matrix revealed several significant relationships among hematological variables and their association with clinical outcomes (see Fig. 2). The strongest positive correlation was observed between the Total Platelet Count and the Platelet to Hematocrit Ratio (r = 0.97), which is expected due to their physiological connection. Similarly, composite indicators such as the Neutrophil to Lymphocyte Ratio and the WBC to Platelet Ratio showed substantial associations with their respective components—Neutrophils (%) (r = 0.73) and Total WBC Count (r = 0.80). This corroborates an inverse relationship b/w Neutrophil to Lymphocyte Ratio and Lymphocytes (%)

(r = -0.87), consistent with the expected immune response pattern of bacterial versus viral infections in which the neutrophil level rises and the lymphocyte count falls. For the diagnostic outcome (result), moderate correlations existed with Neutrophils (%) (r = 0.20), Total WBC count (r = 0.15), and RDW-CV (%) (r = 0.15), and with Lymphocytes (%) being negatively correlated (r = -0.19). These results corroborate the known hematological patterns associated with dengue infection. Conversely, the Severity variable demonstrated merely weak linear relationships over the parameters tested. The strongest associations were for Hemoglobin (r = 0.038) and Lymphocytes (r = 0.049), indicating that whether disease severity is caused by an individual blood parameter identified in this study will require multivariate analysis, as it may be the combination of multiple blood parameters that may lead to disequilibrium.

These results suggest that the tree-based ensembles (Random Forest in this instance) are better at modeling this data than the linear models, which is in line with previous results. In particular, RF's ability to capture nonlinear interactions among CBC features appears advantageous. Nevertheless, the average precision (84.5%) is not as high as reported in some publications [4]. So for now, we used the rest of the results from the test data and just calculated the accuracy; it was 85% where the data was 13,608, which is comparatively a huge number as compared to our dataset. As an example, in RF cases, a higher recall model than precision (meaning fewer false alarms at the expense of more false negatives). The mediocre AUC scores indicate that there are further features or data that might help to create better separation.

4.1 Model Performance and ROC AUC Analysis:

Random Forest got a high accuracy and recall, but ROC AUC = 0.69, which is also low relative to other methods. The reasons for this result can be explained by a few factors:

1. **Class Imbalance:**Even though SMOTE was applied to balance the dataset, the performance of the model in terms of the ROC curve might have been limited due to the initial class imbalance (68% of the instances are dengue-positive, while 32% are negative).
2. **Model Complexity:** Random Forests are resistant to overfitting, but they may not be able to separate the classes due to the complex nature of the intertwined high-correlation features. In our case, blood indices such as platelet count and WBC count may not have enough discriminatory power.
3. **Potential Improvements:**
   Hyperparameter Tuning: Adjusting the parameters of the random forest (e.g., max_depth, min_samples_split) using grid search or random search to better capture the patterns in the data.
   Additional Features: Inclusion of further clinical data (e.g., fever grading, other symptoms) or other biomarkers may help with model discrimination and therefore the AUC.

Alternative Models: Trying out stacking models or ensemble boosting approaches (e.g., LightGBM) might increase the AUC by finding complex interactions between features.

5 Future Work

Future work may extend this study in a number of important ways to enhance the model's performance and practical applicability:

1. **Incorporating Additional Clinical Features:**This study included hematologic features, but other clinical symptoms (e.g., fever, rash, headache), vital signs (e.g., blood pressure, temperature), or other biomarkers (e.g., liver enzymes, cytokine levels) could be incorporated to further improve the predictive performance of the models. These traits may offer a broader perspective of dengue severity and help with early diagnosis.
2. **Longitudinal Data:**Longitudinal data that track the temporal trajectories of blood counts and clinical syndromes for years could further improve the identification of early signals with a high degree of specificity. This would allow models to identify variances in hematological parameters that may present as an early signal of severe dengue, thus enabling prompt intervention.
3. **Multi-Center Data Collection:** A Multicenter study collaborating with diverse locations should be conducted in future research to improve the generalizability of the results. This would allow the model to adjust for variation in the presentation of dengue between regions, making it more generalizable to a diverse population. It would also help illuminate regional epidemiological trends and validate the external validity of the model across different clinical practices.
4. **Advanced Machine Learning Techniques:**Perhaps, using other complex machine learning algorithms may provide better results. For example:
   Deep neural networks (DNNs) or ensemble stacking may find complex interactions in the data that simple models are unable to do.
   Also, **Explainable AI (XAI) methods,** like SHAP or LIME, can be helpful in improving the transparency of the model in terms of its predictions, which could also be important in clinical settings, as gaining clinician trust is highly dependent upon the interpretability of the deep learning model.
5. **Imbalanced Data Strategies:**SMOTE was used in this study; however, other imbalanced data strategies (e.g., ADASYN (Adaptive Synthetic Sampling), cost-sensitive learning) can also be explored. That way, those methods will improve the prediction of the minority class (dengue-negative) and also the major class (dengue-negative).
6. **Real-World Validation and Decision Support Tool Implementation:** The prospective external validation in clinical practice is necessary for further research in order to clarify the practical applicability of the model. For example, the model could be integrated into a decision support system

or a mobile application for real-time use by health professionals. Such tools could enable faster triage and management of dengue against CBC results, especially in a low-resource environment.

7. **Integration with Public Health Surveillance:** Finally, it would enable CDC to trace your model back into the public health surveillance systems for real-time tracking and resource allocation of enzymes for dengue. This type of strategy would allow for a broader epidemiological surveillance and control that could provide for the health of the population.

6 Conclusion

In this study, we systematically evaluate a number of machine learning classifiers on a real-world dengue hematology dataset. In this study, we showed that dengue positivity can be predicted with fair accuracy using models trained on common CBC parameters. The random forest provided the best compromise with respect to sensitivity (91.87% recall) and overall accuracy (84.5%), while neural networks, XGBoost, and SVM achieved similar levels of strong performance. Consistent with clinical intuition as well as prior studies [7], platelet count was identified as the most informative feature. We show how our analysis of tradeoffs between metrics can help for instance, models with higher recall will reduce missed cases of dengue, while higher precision can help to reduce false alarms. Such insights can help the choice of screening approaches according to clinical priorities. In summary, machine learning on conventional hematological data could provide, in the context of dengue diagnosis, a fast and inexpensive decision support tool. If further developed and tested, such models could complement traditional diagnostic workflows, especially in low-resourced areas where there is a high need for early interventions.

References

1. D. R. . P. K. T.S., "Dual level dengue diagnosis using lightweight multilayer perceptron with xai in fog computing environment and rule based inference," *Scientific Reports*, 2025.
2. World Health Organization, "Dengue - global situation," 2024, disease Outbreak News, 30 May 2024.
3. O. F. d. M. L.-J. a. L. A. S. Daniel Cristobal Andrade Girón, William Joel Marín Rodriguez, "Machine learning and deep learning models for dengue diagnosis prediction: A systematic review," *Informatics*, 2025.
4. A. d. J. L. d. A. Claudia Yang Santosa, Suely Tuboia, "A machine learning model to assess potential misdiagnosed dengue hospitalization," *Heliyon*, 2023.
5. J. G. G. O. a. D.-L.-H.-F. Wilson Arrubla-Hoyos, ORCID, "Differential classification of dengue, zika, and chikungunya using machine learning—random forest and decision tree techniques," *MDPI*, 2024.
6. A. A. Rao, A. Sivakumar, S. Vishnubhatla, V. Kumar, N. Simon, P. Makkar, M. Srinivasan, A. Singh, and S. Agarwal, "Dengue fever: Prognostic insights from a complete blood count," *J. Family Med. Prim. Care*, 2020.

7. S. Dasgupta, S. Das, and D. Chakraborty, "Machine learning-based mathematical equations for dengue positivity detection using elementary laboratory parameters," *J. Family Med. Prim. Care*, 2025.
8. T. Chilakala and R. Garladinne, "Artificial intelligence-based early detection of dengue using cbc data," *J. Inf. Syst. Eng. Manag.*, 2025.
9. A. H. H. H. J. A. Z. J. A. Ariba Qaiser, Sobia Manzoor, "Support vector machine outperforms other machine learning models in early diagnosis of dengue using routine clinical data," *National Library of Medicine*, 2025.
10. S. C. M. S. A. T. S. D. D. F. V. S. B. J. S. L. F. R. P. B. Bianca Conrad Bohm, Fernando Elias de Melo Borges, "Utilization of machine learning for dengue case screening," *National Library of Medicine*, 2024.
11. M. Bairy, K. Chitthur, B. Sundararao, R. R. Hire, V. Aparna, S. D. Chandrashekar, and P. V. Magadi, "Machine learning-based detection of dengue from blood smear images utilizing platelet and lymphocyte characteristics," *Diagnostics (Basel)*, 2023.
12. B. L. R. R. P. F. H. B. D. A. C. G. A. V. P. L. A. N. L. J. B. Carmen Alicia Ruiz Valdez, Olga María Alejo Martínez, "Effectiveness of a diagnostic algorithm for dengue based on an artificial neural network," *National Library of Medicine*, 2024.
13. . N. S. . O. S. B. . Hilda Mayrose 1ORCID, G. Muralidhar Bairy 1, "Machine learning-based detection of dengue from blood smear images utilizing platelet and lymphocyte characteristics," *MDPI*, 2023.
14. A. H. U. R. S. Samrat Kumar Dey, Md Mahbubur Rahman, "Prediction of dengue incidents using hospitalized patients, metrological and socio-economic data in bangladesh: A machine learning approach," *National Library of Medicine*, 2022.
15. M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: Review of a decade of research," *Artificial Intelligence Review*, 2024.