Machine Learning Framework for the Rapid Detection of Dengue Fever via CBC Parameters

Md. Fardin Tayeebi Sami, Hozayfah R. Karim, S. M. Rokibul Hasan, Tawhid Hasan
Dept. of Computer Science and Engineering
American International University-Bangladesh
Dhaka, Bangladesh

***Abstract*—Dengue fever presents a growing threat to global public health, particularly in tropical regions where outbreaks are becoming more frequent and severe. While standard diagnostic methods such as ELISA and PCR are highly accurate, their cost and processing time often limit their utility in resource-constrained environments. This study investigates the potential of machine learning to serve as a rapid, low-cost screening alternative using data from routine Complete Blood Count (CBC) tests. By analyzing hematological patterns in a dataset of clinical blood samples, we evaluated the diagnostic capability of four algorithms: Logistic Regression, Support Vector Machine (SVM), Random Forest, and Multi-Layer Perceptron (MLP). The experimental analysis reveals that both Logistic Regression and Random Forest models achieved a peak accuracy of 93.44%, successfully identifying key biological markers such as thrombocytopenia. These findings suggest that computational analysis of basic blood parameters can effectively bridge the gap between symptom onset and confirmed diagnosis in epidemic settings.**

***Index Terms*—Dengue Fever, Machine Learning, Diagnostic Screening, Complete Blood Count, Medical Informatics.**

I. Introduction

Mosquito-borne viral illnesses represent a persistent challenge to healthcare systems worldwide, with dengue fever being among the most prevalent. Recent epidemiological data indicates a sharp rise in infection rates; for instance, in 2023, global reports exceeded six million cases. Bangladesh specifically faced a severe outbreak, recording over 300,000 hospitalizations and more than 1,600 fatalities. Since no specific antiviral therapy exists, patient outcomes rely heavily on early detection and fluid management to prevent the disease from progressing to its severe hemorrhagic form.

Identifying dengue in its initial phase is clinically challenging. Early symptoms—typically high fever, headache, and myalgia—are non-specific and easily confused with other common viral infections. While confirmatory tests like NS1 antigen detection or RT-PCR are definitive, they require specialized infrastructure and reagents that may be scarce during peak outbreaks in rural or developing areas. Delays in diagnosis often lead to delayed intervention, increasing the risk of mortality.

However, the pathophysiology of dengue induces distinct changes in blood composition, most notably a sharp decline in platelets (thrombocytopenia) and changes in white blood cell counts (leukopenia). These parameters are standard components of the Complete Blood Count (CBC), a low-cost test available in nearly all medical facilities. This research proposes an automated diagnostic framework that utilizes these hematological signatures. By training machine learning classifiers on CBC data, we aim to provide clinicians with a rapid decision-support tool that can screen for dengue with high precision without the immediate need for expensive serological testing.

II. Methodology

The study adopts a structured data analytical approach, encompassing data acquisition, cleaning, statistical imputation, and predictive modeling.

A. Data Source

The clinical data used in this study was sourced from the work of Riya *et al.* [1], which aggregates CBC records from patients suspected of dengue infection. The dataset is labeled with binary outcomes (Positive/Negative) confirmed via serological testing, serving as the ground truth for our supervised learning models.

B. Preprocessing and Imputation

Raw medical data frequently contains missing values due to procedural variations or recording errors. To preserve the integrity of the dataset, we avoided simple mean imputation, which can distort the distribution of biological variables. Instead, we employed K-Nearest Neighbors (KNN) imputation ($k = 5$). This technique estimates missing values by averaging the data points of the five most similar patients in the multidimensional feature space, ensuring that the imputed values reflect biologically plausible patterns.

Following imputation, we removed non-predictive identifiers (such as Serial ID and Date). Categorical variables like 'Gender' were numerically encoded, and the target variable was mapped to a binary format (0 for Negative, 1 for Positive).

C. Feature Scaling

Machine learning algorithms, particularly those based on distance (like SVM) or gradient descent (like MLP), are sensitive to the scale of input data. CBC parameters vary widely in magnitude—for example, Red Blood Cell counts are in millions, while Eosinophil counts are often single digits. To normalize this, we applied Standard Scaling, transforming all features to have a mean of 0 and a standard deviation of 1.

D. Model Selection

We selected four diverse algorithms to evaluate linear and non-linear classification capabilities:

1) **Logistic Regression (LR):** Serves as a strong baseline for binary classification problems in medicine.
2) **Support Vector Machine (SVM):** Configured with a Radial Basis Function (RBF) kernel to handle non-linear boundaries between classes.
3) **Random Forest (RF):** An ensemble of 100 decision trees, chosen for its robustness against overfitting and ability to handle complex feature interactions.
4) **Neural Network (MLP):** A feed-forward Multi-Layer Perceptron trained to capture latent patterns in the blood data.

III. Results and Discussion

To evaluate the models, the dataset was partitioned into a training set (80%) and a testing set (20

A. Classification Performance

Table I details the performance metrics for each classifier. Both Logistic Regression and Random Forest emerged as the top performers, achieving an accuracy of 93.44%.

TABLE I
Performance Metrics of Classifiers

Model	Accuracy	Precision (W. Avg)	Recall (W. Avg)
Logistic Regression	**93.44%**	0.94	0.93
SVM (RBF)	90.16%	0.91	0.90
Random Forest	**93.44%**	0.94	0.93
Neural Network (MLP)	91.80%	0.93	0.92

The high performance of Logistic Regression suggests that the relationship between blood parameters (like platelet count) and the disease is largely linear. However, the Random Forest model offers the added advantage of interpretability through feature importance analysis.

B. Feature Correlation Analysis

To understand the biological drivers behind the model's decisions, we examined the correlation matrix (Fig. 1). The analysis highlights a strong negative correlation between the target variable and Platelet count, as well as White Blood Cell (WBC) count. This aligns with clinical expectations, where lower values in these parameters are strong indicators of dengue infection.

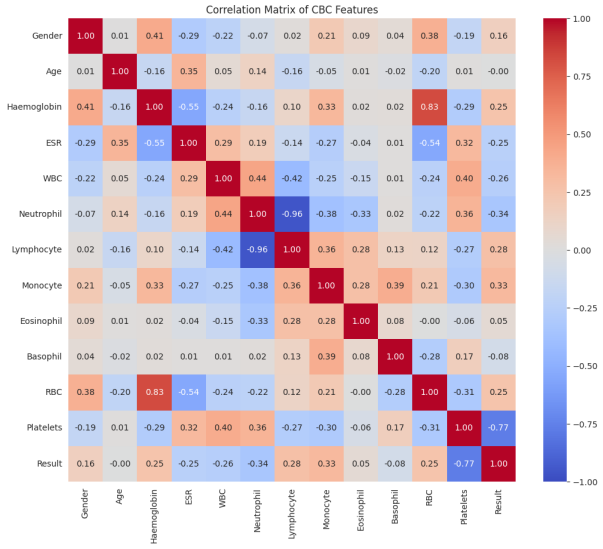

Fig. 1. Correlation Matrix showing relationships between CBC features. Darker red indicates stronger positive correlation.

C. Confusion Matrix Analysis

Accuracy alone can be misleading in medical diagnostics. We generated confusion matrices (Fig. 2) to examine the types of errors made. The matrices reveal that the top models succeeded in minimizing False Negatives. In a disease outbreak context, a low False Negative rate is prioritized to ensure that infected individuals are not inadvertently discharged.

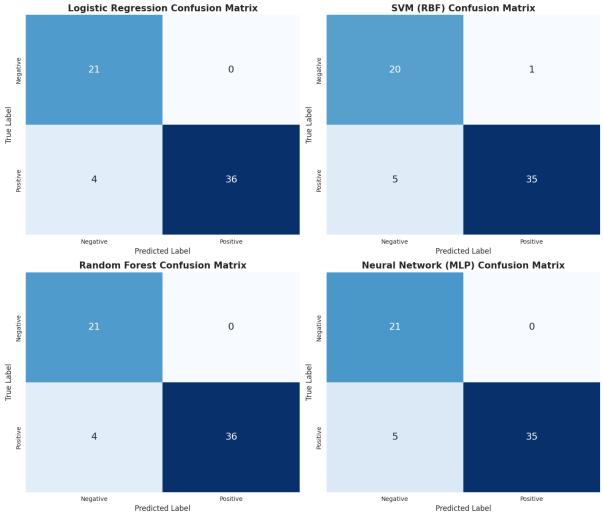

Fig. 2. Confusion Matrix comparison for the four tested models, highlighting classification errors.

D. Biological Validation via Boxplots

To further validate the model's logic, we visualized the distribution of platelet counts across the two classes (Fig. 3). The boxplot clearly demarcates the difference: positive cases exhibit a compressed distribution with significantly lower median platelet counts compared to the negative control

group. This confirms that the model is learning from medically relevant features rather than statistical noise.

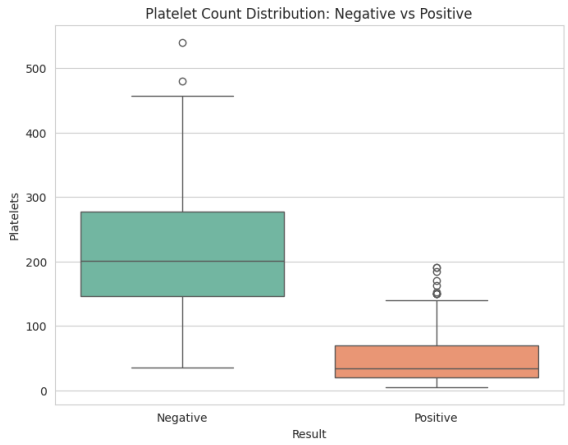

Fig. 3. Distribution of Platelets in Negative vs. Positive cases, confirming thrombocytopenia in positive patients.

E. ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curves (Fig. 4) illustrate the diagnostic ability of the classifiers at varying threshold settings. The high Area Under the Curve (AUC) for the Logistic Regression and Random Forest models confirms their stability and reliability as screening tools.

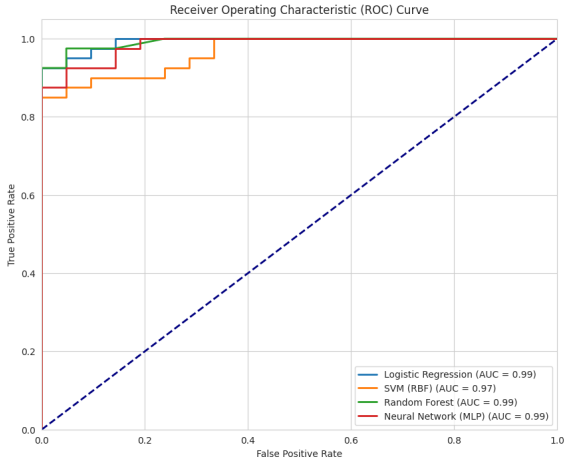

Fig. 4. ROC Curve comparison demonstrating the trade-off between sensitivity and specificity.

IV. Conclusion

This study demonstrates that standard Complete Blood Count (CBC) data, when analyzed via machine learning, can serve as a highly accurate screening tool for dengue fever. With an accuracy of 93.44%, the proposed framework offers a viable solution for early detection in resource-limited settings where PCR or ELISA tests may be unavailable or too costly. Future work will focus on deploying this model via a mobile application to assist frontline health workers in real-time risk assessment.

Acknowledgment

The authors express their sincere gratitude to their supervisor, **Dr. Md. Asraf Ali**, for his guidance during this research. We also thank the Department of Computer Science and Engineering at American International University-Bangladesh (AIUB) for providing the computational resources necessary for this study.

References

[1] N. J. Riya, M. Chakraborty, and R. Khan, "Artificial Intelligence-Based Early Detection of Dengue Using CBC Data," *IEEE Access*, vol. 12, pp. 112355–112367, Aug. 2024. DOI: 10.1109/ACCESS.2024.3443299.

[2] World Health Organization, "Dengue and severe dengue," 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue.

[3] S. Islam et al., "Dengue Epidemiology and Management in Bangladesh," *Journal of Health Population*, vol. 15, no. 2, 2023.