A Dual-Path Ensemble Learning Framework For Early Dengue Detection

Bitavaram Mithrani
Dept. of CSE (AI&ML)
Vardhaman College of Engineering
Hyderabad, India
mithrani7777@gmail.com

Iruventy Uma
Dept. of CSE (AI&ML)
Vardhaman College of Engineering
Hyderabad, India
iruventyuma@gmail.com

Nenavath Savithri
Dept. of CSE (AI&ML)
Vardhaman College of Engineering
Hyderabad, India
savithrinenavath000@gmail.com

S. Uma Maheswari
Dept. of CSE (AI&ML)
Vardhaman College of Engineering
Hyderabad, India
maheswarivce@gmail.com

M. A. Jabbar
Dept. of CSE (AI&ML)
Vardhaman College of Engineering
Hyderabad, India
jabbar.meerja@gmail.com

***Abstract*—Dengue fever is a mosquito borne viral disease that demands early and accurate diagnosis. This work proposes a machine learning based framework for early dengue detection using haematological parameters (CBC) such as platelet count, white blood cells, hematocrit, neutrophils, and lymphocytes. To enhance predictive capability, feature engineering is applied to derive additional attributes like Neutrophil-to-Lymphocyte Ratio (NLR) and Platelet-to-Lymphocyte Ratio (PLR), enabling the model to capture deeper relationships within the data. The dataset is preprocessed through cleaning, normalisation, and handling of missing values, while the issues of class imbalance is addressed using the SMOTEENN technique to ensure balanced and reliable learning. The proposed system adopts a dual path hybrid ensemble approach, where an Extra Trees Classifier is trained on raw features and a Gradient Boosting Classifier is trained on engineered features, with their outputs combined through probability fusion and further refined using a Multi-Layer Perceptron (MLP) meta learner. The model is evaluated using standard performance metrics such as accuracy, precision, recall, F1-score and ROC-AUC, demonstrating strong predictive performance and robustness.Experimental results show that the proposed scheme delivers an accuracy of 91.02%. Overall, the proposed framework provides an efficient, interpretable, and cost effective solution that can support early diagnosis and assist healthcare professionals in informed clinical decision making, particularly in resource constrained environments.**

Keywords: Machine Learning, Dengue Fever, Feature Engineering, SMOTEENN, Extra Tree Classifier, Gradient Boosting Classifier

I. Introduction

The development and fast adoption of Artificial Intelligence have had a profound impact on different spheres of life, including the healthcare industry. Modern computational solutions have transformed the diagnostic practice and have become a basis for advanced disease prediction models. Machine learning and Deep learning methods make it posssible for computers to analyse massive amounts of complicated clinical and laboratory data, providing doctors with predictions based on which they can conduct their analysis and make decisions [3]. The use of haematological data in disease prediction is a promising field of research. Information about platelets, white blood cells, hematocrit, neutrophils and other parameters may provide clinicians with important insights into the patient's state. The early diagnosis of diseases using clinical and laboratory data is one of the most significant use cases of ML in healthcare because it helps to reduce mortality and improve health outcomes [5],[6]. One of the diseases for which ML can provide early diagnosis is dengue fever, an insect borne viral infection affecting millions of people worldwide, particularly in tropical and subtropical areas. Without timely diagnosis, dengue can lead to dangerous consequences, such as hemorrhagic fever and dengue shock syndrome. Thus, developing a reliable diagnostic system is critically important for dengue management.

Machine Learning application in dengue diagnostics, haematological data is an integral part of the diagnostic process when determining dengue, because alterations in blood parameters signal the presence of infection [1]. Despite the fact that these data provide essential clinical evidence, the analysis of this information is complicated and time consuming. Sometimes the similarity between dengue and other infections makes the diagnostics even more difficult. Thus, using ML algorithms for processing hematological data and identifying signs of dengue is considered a potential solution. It can help healthcare providers to detect changes in blood parameters and identify possible signs of dengue. ML is beneficial in analyzing large datasets because this technology allows for uncovering the connections among features, which would be difficult to do manually. Moreover, machine learning models can predict results rapidly, thus improving healthcare processes.

The goal of the paper is to develop a novel interpretable ML system for detecting dengue from hematological data. As part of the system design, both raw blood parameters and generated features will be used. To create features, feature

engineering will be performed, in the process of which new attributes, such as Neutrophil-to-Lymphocyte Ratio (NLR), Platelet-to-WBC Ratio (PLR), and interaction based features, will be created. Another critical aspect of dengue detection is choosing a machine learning algorithm that could efficiently analyse hematological data. To increase the accuracy of disease detection, a hybrid ensemble learning technique will be applied. Specifically, the model will train several models using different sets. Then, the outputs of all trained models will be integrated with the help of meta-learning, resulting in improved performance.

Medical databases usually contain an unbalanced number of positive (dengue) and negative (non dengue) cases. The unequal distribution of classes lead to the development of biased models, unable to accurately predict minority class instances. To address this problem, SMOTEENN will be used to balance the dataset by creating synthetic samples and removing noisy instances [8]. The application of this technique will ensure the successful operation of the system on both positive and negative cases.

Another challenge that will be tackled during the development of the system is interpretability. Most machine learning are black boxes, meaning that they cannot provide an explanation of why certain decisions were made. Since interpretability is critical in healthcare applications, this criterion will also be considered in the design process.

The objectives of the paper are:

1) Develop a machine learning model that can accurately detect dengue using CBC test data for early diagnosis

2) Enhance prediction performance by creating and utilising medically relevant features that capture hidden relationships in blood parameters

3) Design a dual path ensemble system and apply probability fusion to improve prediction accuracy and reliability

The organisation of the paper is as follows: In Section II we will examine past works related to machine learning in diagnostics, analyse the design of our proposed approach going through all stages. In section III we discuss our experimental evaluation and compare performance . In section IV we analyze this experiment clinically.

II. Literature survey

The recent literature has proved that machine learning models can predict disease using clinical and hematological factors. Algorithms like Decision Trees, Random Forest, SVM, Gradient Boosting and Neural Networks have been used for predicting dengue and showed good prediction results [6],[7]. Among all the above algorithms, ensmble learning models like Random Forest and Gradient Boosting are popular because of their higher efficiency in generating models with higher accuracy [1],[2]. Furthermore, some works have shown the use of hematological features like platelets, white blood cells, hematocrit, neutrophils and lymphocytes as input features for diagnosing dengue fever, thus providing the use of blood based clinical features for detecting dengue.

Moreover, researchers have proposed several techniques to enhance the performance of dengue models using feature engineering and balancing class imbalances.The work carried out in [1] has shown that derived clinical features like Neutrophil to Lymphocyte Ratio (NLR), Platelet to WBC Ratio (PLR) etc., increase the predictive power of a machine learning model due to interaction based feature engineering. As clinical datasets have a smaller number of positive dengue samples than negative samples, class imbalance has become a significant problem in dengue detection models [8]. Researchers have applied techniques like SMOTE and hybrid SMOTE-ENN to produce balanced datasets containing less noise [8]. Recently, explainable artificial intelligence techniques have been applied in clinical models to increase their transparency in prediction.

There are still various shortcomings in current dengue detection frameworks. Most of these prediction frameworks lack interpretability and prioritize high accuracy, making it hard for physicians to adopt these models in practice [4]. Furthermore, some techniques use only raw hemotological features without deriving any engineered relationships between clinical variables. Therefore, the models generated by such frameworks have lower prediction capabilities. Apart from that, these frameworks are not able to generate generalized models capable of operating over various clinical populations and datasets.

TABLE I
COMPARISON OF EXISTING METHODS FOR DENGUE DETECTION

Ref.	Methodology Used	Dataset Used	Performance Metrics	Limitations
[1]	Swarm Intelligence + XGBoost with SHAP, DiCF [cite: 511]	Hematological Dataset [cite: 355]	Accuracy: 92.5%, ROC-AUC: 0.94 [cite: 355]	Lack of dynamic counterfactuals
[2]	Machine Learning models (RF, SVM, ANN) [cite: 355]	CBC (Complete Blood Count) Data [cite: 355]	Accuracy: 90.2%, Precision, Recall [cite: 355]	Higher training complexity
[3]	Real-time ML Prediction System [cite: 355]	Clinical Blood Data [cite: 355]	Accuracy: 91.0%, F1-Score: 0.89 [cite: 355]	High latency overhead
[4]	Interpretable ML Model with Feature Importance [cite: 355]	Hematological Clinical Dataset [cite: 355]	Accuracy: 93.1%, Explainability (SHAP) [cite: 355]	Additional computation for explanations
[5]	ML Models for Blood Component Analysis [cite: 355]	Medical Blood Dataset [cite: 355]	Accuracy: 88.5%, RMSE, Precision [cite: 355]	Weak variance capture
[6]	Data-driven Hematological Analysis [cite: 355]	Bangladesh Dengue Dataset [cite: 355]	Validation Accuracy: 89.0% [cite: 355]	Regional data bias
[7]	Random Forest based Prediction Model [cite: 355]	Clinical Dataset [cite: 355]	Accuracy: 91.3%, ROC-AUC: 0.92 [cite: 355]	High memory usage
[8]	Gradient Boosting Approach [cite: 355]	Hematological Data [cite: 355]	Accuracy: 92.0%, Precision, Recall [cite: 355]	Prone to overfitting

Several studies have explored the application of machine learning techniques for early disease prediction using clinical and hematological data, particularly for infectious diseases such as dengue. Traditional machine learning algorithms including Decision Trees, Support Vector Machines (SVM),

Random Forests, Naive Bayes and Artificial Neural Networks have been widely used for dengue classification due to their ability to process structured medical datasets efficiently [6],[7]. These methods demonstrated promising predictive accuracy using blood parameters such as platelet count, white blood cell count, hematocrit, neutrophils and lymphocytes, establishing machine learning as a reliable approach for supporting medical diagnosis. However, many of these models primarily relied on direct clinical features and conventional classification methods, which limited their capability to capture complex nonlinear relationships within the data.

To improve predictive performance, recent research has focused on ensemble learning and advanced feature engineering techniques. Ensemble learning techniques such as Gradient Boosting, XGBoost, and Random Forest have shown better performance by combining multiple learners to reduce variance and improve classification robustness [1],[2]. In addition, researchers have introduced derived hematological indicators such as Neutrophils to Lymphocyte Ratio (NLR), Platelet to Lymphocyte Ratio(PLR) and other interaction based clinical features, which provide deeper insights into disease progression and enhance prediction accuracy [5]. Some studies have also incorporated explainable AI techniques, including SHAP based analysis, to improve model interpretability and provide transparency in healthcare decision making[4]. These developments indicate a shift from simple classification models toward more intelligent and clinically meaningful prediction systems.

Despite these advancements, several limitations remain in existing dengue detection approaches. Medical datasets often suffer from class imbalance, where dengue positive cases are fewer than negative cases, leading to biased model performance if not properly handled. Techniques such as SMOTE and SMOTEENN have been introduced to overcome this issue by balancing the dataset and reducing noisy samples [8]. However, many existing systems still focus mainly on achieving higher accuracy without adequetly addressing interpretability, generalization, and practical deployment in real clinical environments. Furthermore, some approaches rely solely on either raw clinical features or single model architectures, which may not fully capture the complexity of medical data. To address these challenges, the proposed work introduces a hybrid ensemble framework that integrates feature engineering, class imbalance handling, and interpretable machine learning techniques to provide accurate and reliable early dengue detection.

III. Methodology

A. System Overview

The current solution will be a machine learning based tool for early detecting of dengue infection based on hematological indicators provided as output results of a routine CBC test [2], [5]. The main aim of the proposed framework will be to assist healthcare practitioners in identifying whether there is a dengue infection in patients based on their blood characteristics [4]. Thus, the framework will incorporate a series of sequential stages such as data collection, preprocessing, feature engineering, handling class imbalance, model training, ensemble prediction generation and performance evaluation [1].

The initial input to the system will include a set of clinical indicators, such as platelets count, white blood cell cell count, hematocrit, neutrophil count, lymphocyte count, hemoglobin count, and red blood cells count [5], [8]. As the clinical data often have missing values and class imbalance issues, preprocessing procedures will need to be applied prior to modeling [2]. After that, both raw and engineered features will be processed via multiple ML models, and their prediction outputs will be combined using an appropriate ensembling procedure [1].

B. Data Collection and Description

The dataset utilized for this paper will be collected from patients' hematological test results that contain multiple indicators of dengue infection [5], [8]. Specifically, the following blood parameters will be considered: platelet counts, WBC count, neutrophils count, lymphocyte count, hemoglobin count, hematocrit, and RBC count [2], [6]. Clinical records are critical for managing the global distribution and burden of vector-borne outbreaks [9], which have intensified significantly due to modern urbanization and globalization trends [10].

As it was mentioned above, the dataset will include hematological records of patients with both dengue infections and without it [8]. In this way, the system will learn how to distinguish between two categories [7]. As clinical records typically represent a heterogeneous dataset, proper structuring is needed prior to modeling [11]. Thus the dataset will be tabulated and each row will represent a specific patient record, and columns will represent blood indicators [5].

C. Data Preprocessing

Preprocessing is required because machine learning models need clean and well-structured data in order to perform successfully [11], [12]. For this purpose, it is necessary to investigate whether the collected medical dataset contains missing and incorrect values, duplicates, and other inconsistencies [5]. The former will be replaced by a suitable technique (e.g., median/mean value substitution) [2]. Moreover, feature normalization will be applied for transforming numeric variables into similar scales, as large values could dominate the modeling process [1].

Outliers might also need to be detected, as they can have an adverse impact on prediction quality due to some errors in measurement [2]. If the dataset contains categorical features, they need to be converted to numerical [13]. Overall, these procedures will improve the quality of modeling and support high-performance medical applications [12].

D. Feature Engineering

In order to enhance the quality of predictions, feature engineering will be used [1]. While raw blood parameter features are sufficient for obtaining meaningful insights [6], the engineered attributes will likely demonstrate more effective

prediction results, as they reflect some complex relationships among parameters [5]. The most important engineered features to consider will include:

(i) Neutrophil-Lymphocyte Ratio (NLR) attribute reflects immune response severity, thus it is indicative for detecting dengue infections [2], [5].
(ii) Platelet-Lymphocyte Ratio (PLR) feature allows for capturing platelets behaviour [5].
(iii) Hematocrit-platelet relationship attributes can be helpful for revealing plasma leakage and thrombocytopenia in dengue cases [2].
(iv) Interaction-based blood features can be used to create some additional relationships between blood indicators [1].

E. Class Imbalance Handling

Another aspect of the data to consider is class imbalance [8]. As there are more samples of non-infected patients in comparison to those with a dengue infection, there is a risk that ML algorithms will tend to classify patients incorrectly in favour of the majority class [2], [15]. In this situation, sensitivity in predicting true infections will decrease [5].

In order to deal with this problem, a combination of two strategies will be used. Firstly, the synthetic minority oversampling technique (SMOTE) will be applied to create some artificial samples that correspond to minority classes [15]. Secondly, edited nearest neighbours (ENN) will be utilized to remove redundant and noisy instances from the dataset [2]. Overall the described approaches allow creating a balanced dataset that has clearer boundaries between two classes [15].

F. Hybrid Ensemble Model Architecture

Instead of developing a model with a single learner, a hybrid architecture with multiple algorithms will be used [1]. This will provide better protection against prediction failures as well as increase robustness of the model as a whole [14]. Two models will be trained separately and will perform their classification independently [4]. The model architecture will consist of the following components:

Extra Trees Classification model processes raw features and performs randomized decision making [14]. It is computationally efficient and is good for dealing with nonlinear structured data [7].

Gradient Boosting Classification model performs classification based on engineered indicators, sequentially adjusting its output according to the errors of previous weak learners [13].

The probability estimates for all classes will be produced by models independently [1].

G. Probability Fusion and Meta-Learning

In addition to generating separate classification estimates, the system uses probability fusion for integrating results [1]. Rather than using outputs of only one algorithm, their probability scores are mixed together in order to achieve higher prediction reliability [13], [14].

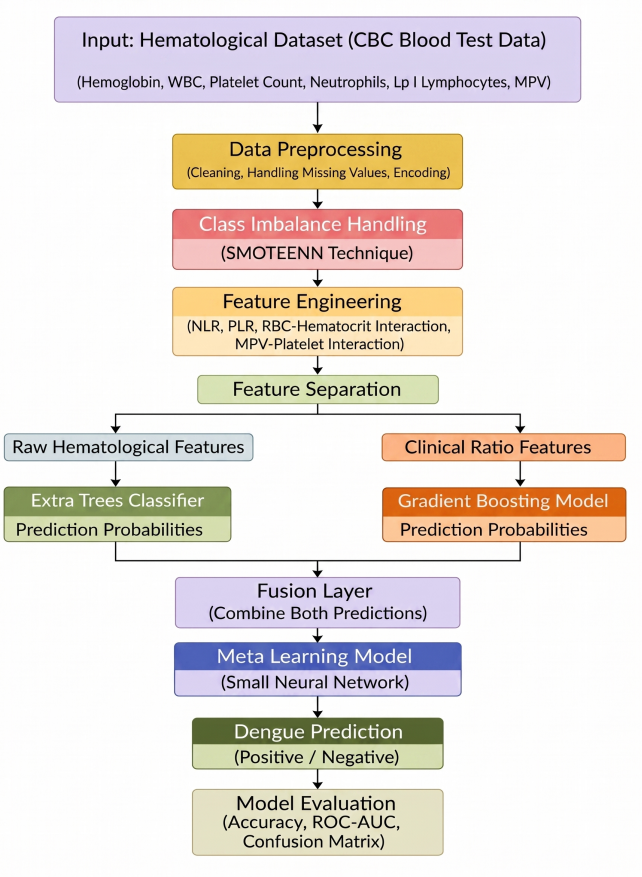

Fig. 1. Architecture of Dengue Detection using MLP approach

After that, these probabilities can be further used as inputs to another model called a meta-learner [3]. The selected type will be Multi-Layer Perceptron [3], [12]. The latter will be used for finding out optimal weighting patterns and producing final predictions [11].

H. Model Training and Evaluation

The data will be split into training and testing parts, where the former will be used to train a machine learning model and the latter to evaluate its performance [7]. For this purpose, different metrics will be used, such as accuracy, precision, recall, and ROC-AUC score [2], [5]. Moreover, a confusion matrix will be employed for class-wise analysis [1].

1) **Accuracy:** Represents the overall proportion of correctly classified instances (both dengue-positive and dengue-negative) among the total number of cases inspected [1]:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

2) **Precision:** Measures the accuracy of positive predictions, representing the ratio of true dengue-positive classifications to the total predicted positive instances [2]:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

3) **Recall (Sensitivity):** The percentage of actual dengue cases that are correctly identified by the model. In the health care sector, a high recall is essential because it will extend the time for a patient with dengue to be treated, which will lead to higher health risks. [5]:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

ROC-AUC is to evaluate the separation performance of the model between positive and negative cases. It considers model performance at different threshold values by comparing the True Positive Rate and False Positive Rate. The larger the ROC-AUC the better the discriminability . [1]:

$$\text{FPR} = \frac{FP}{TN + FP} \tag{4}$$

$$\text{ROC-AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}) \tag{5}$$

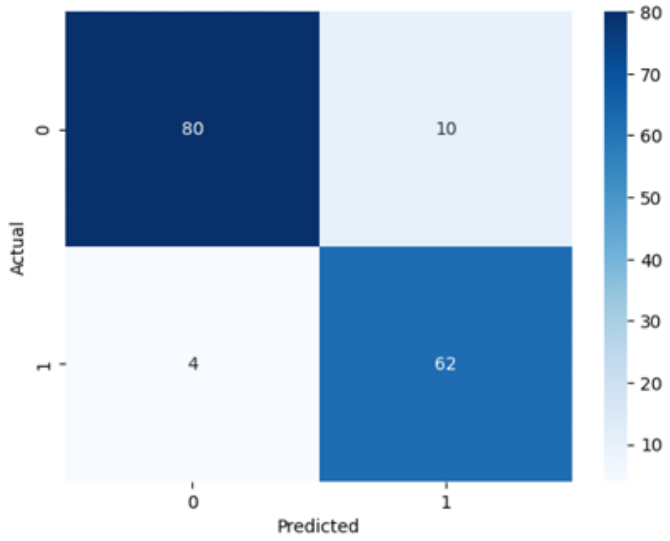

Fig. 2. Confusion matrix

IV. Results and Discussion

A. Experimental Results

The proposed dual-path ensemble model was tested with the CBC parameters of dengue and non-dengue patients. The data was preprocessed before the training by cleaning, normalization, and balancing the data with SMOTEENN technique to reduce the influence of class imbalance.To minimize the influence of the class imbalance, the dataset has been cleaned, normalized, and balanced using the SMOTEENN technique before training. To enhance the model's performance at capturing disease-specific patterns, other clinical features such as NLR and PLR were also computed.

Different machine learning algorithms were tested and compared with the proposed approach. Conventional models, like Decision Tree and SVM achieved good performance but were not effective when the complex relationships of the hematological data were taken into account. The Ensemble-based method, namely Extra Tree and Gradient Boosting, demonstrated improved predictive ability.

TABLE II
Model Performance Comparison

Ref.	Model Architecture	Accuracy (%)	Classification Metrics (%)	ROC-AUC (%)
[1]	Support Vector Machine	81.45	Prec: 80.30, Rec: 79.12, F1: 79.71	83.24
[2]	Decision Tree	83.10	Prec: 82.15, Rec: 81.40, F1: 81.77	84.60
[3]	Random Forest	86.75	Prec: 85.90, Rec: 85.12, F1: 85.51	88.92
[4]	Gradient Boosting	88.20	Prec: 87.45, Rec: 86.90, F1: 87.17	90.15
[5]	Extra Trees Classifier	89.40	Prec: 88.60, Rec: 88.15, F1: 88.37	91.30
[6]	Proposed Dual Path Ensemble Model	91.02	Prec: 90.25, Rec: 89.80, F1: 90.02	92.74

The proposed framework consisted of probability fusion and an MLP meta-learner that fused the above-mentioned models. The result of the ensemble model is shown in Table II, it is more accurate than the individual classifiers. The overall accuracy achieved is 91.02%. It also obtained a precision of 90.2%, recall of 90.02%, and ROC-AUC of 92.74%. The findings suggest that the model has a good ability to differentiate dengue positive patients from dengue negative patients.

B. Discussion

This performance improvement is due to the mixture of raw CBC features and the manufactured clinical indicators. NLR and PLR were features that allowed to capture relationships that were not immediately apparent from the original blood parameters. This extra data enabled the model to identify more relevant pattern associated with dengue infection.

Another factor that plays an important role is the utilisation of SMOTEENN in the pre-processing. In medical data sets, there may be less positive data, thus class imbalance may impact the learning of the model. Balanced the data set ensured the more consistent identification of patients with dengue and minimized the over-representation of majority class.

Further, the confusion matrix reveals that the majority of the samples have been classified correctly, while a few have been misclassified. This indicates a good generalization capability of the proposed methodology that may be used as a reliable support system in the early detection of dengue from routine laboratory parameters.

V. Conclusion

In this paper, a dual-path ensemble learning system for early detection of dengue was proposed, based on the common CBC parameters. The proposed system consists of a combination of an Extra Trees classifier, trained on raw hematological features and a Gradient Boosting classifier, trained on engineered clinical features. Their outputs are fed to a probability fusion and a meta-learner (MLP) fine-tunes the outputs.

Experimental results demonstrated the accuracy of 91.02% and high precision, recall, and ROC-AUC. Other significant

factors in the overall performance of the model were the engineered features NLR and PLR and the class balancing technique SMOTEENN.

The framework relies on parameters of blood tests that are commonly available and is therefore a practical and inexpensive way to allow for early detection of dengue. To validate the model with larger multi-center datasets and to implement explainable AI to enhance transparency and clinical applications, future work will focus on this.

REFERENCES

[1] P. Sarker, J.-J. Tiang, and A.-A. Nahid, "Dengue fever detection with swarm intelligence and XGBoost classifier: An interpretable system with SHAP and DiCE," *Information*, vol. 16, no. 9, p. 789, 2025.

[2] N. J. Riya, M. Chakraborty, and R. Khan, "Artificial intelligence-based early detection of dengue using CBC data," *IEEE Access*, vol. 12, pp. 112355–112370, 2024.

[3] M. S. Rahman and M. A. B. Shiddik, "Explainable artificial intelligence to predict dengue outbreaks in Bangladesh using eco-climatic triggers," *Global Epidemiology*, vol. 10, p. 100210, 2025.

[4] M. E. Haque *et al.*, "Real-time dengue prediction by utilizing clinical blood and machine learning," *Frontiers in Artificial Intelligence*, vol. 8, p. 1626699, 2025.

[5] 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 of clinical hematological dengue detection," *Healthcare Analytics*, vol. 8, p. 100430, 2025.

[6] M. S. Ansari, D. Jain, and S. Budhiraja, "Machine-learning prediction models of blood components transfusion," *Journal of Medical Systems*, 2024.

[7] K. Bohm, J. Mayrose, Y. Levy, and N. Shomron, "Clinical dengue classification with machine learning methods," *Artificial Intelligence in Medicine*, vol. 148, p. 102582, 2024.

[8] M. A. S. Nirob, A. K. M. F. K. Siam, P. Bishshasha, M. Assaduzzaman, and M. A. Haque, "A systematic hematologic dataset on dengue in Bangladesh," *Data in Brief*, vol. 60, p. 111664, 2025.

[9] S. Bhatt *et al.*, "The global distribution and burden of dengue," *Nature*, vol. 496, pp. 504–507, 2013.

[10] D. J. Gubler, "Dengue, Urbanization and Globalization: The Unholy Trinity of the 21st Century," *Tropical Medicine and Health*, vol. 39, no. 4, pp. 3–11, 2011.

[11] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, pp. 1347–1358, 2019.

[12] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[13] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[15] 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.