A Study on Hybrid Ensemble Approaches for Effective Dengue Detection and Classification

<u>Supervised By</u>

Afsana Begum

Assistant Professor & Coordinator M. Sc

Department of Software Engineering

Daffodil International University

<u>Submitted By</u>

Abdullah Al Nayem

ID:213-35-767

Department of Software Engineering

Daffodil International University

This thesis report has been submitted in fulfilment of the requirements for the Degree of Bachelor of Science in Software Engineering.

APPROVAL

APPROVAL

This thesis titled on **"A Study on Hybrid Ensemble Approaches for Effective Dengue Detection and Classification,"** submitted by **Abdullah Al Nayem (ID: 213-35-767)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

______________________________ **Chairman**

Dr. Imran Mahmud
Professor & Head
Department of Software Engineering
Faculty of Science and Information Technology Daffodil International University

______________________________ **Internal Examiner 1**

Afsana Begum
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

______________________________ **Internal Examiner 2**

Md. Shohel Arman
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

______________________________ **Internal Examiner 3**

Nadira Islam
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

______________________________ **External Examiner**

Md Manowarul Islam
Professor
Department of Computer Science and Engineering
Jagannath University, Bangladesh

A Study on Hybrid Ensemble Approaches for Effective Dengue Detection and Classification

Abdullah Al Nayem

ID:213-35-767

Bachelor of Science

DAFFODIL INTERNATIONAL UNIVERSITY

SUPERVISOR’S DECLARATION

I hereby declare that I have reviewed this thesis entitled **"A Study on Hybrid Ensemble Approaches for Effective Dengue Detection and Classification "**, and in my opinion, it is adequate in terms of scope and quality for the award of the degree of Bachelor of Science in Software Engineering.

(Supervisor’s Signature)

Full Name : Afsana Begum

Position : Assistant Professor & Coordinator M. Sc

Date : 25 December 2025

STUDENT'S DECLARATION

I confirm that the piece in this thesis is based on my own writing with the exception of quotation and reference that have been discussed. I also confirm that it was not previously and concurrently registered at Daffodil International University or other institutions at any other degree.

Nayem

(Student's Signature)

Full Name : Abdullah Al Nayem

ID Number : 213-35-767

Date : 25 December 2025

A Study on Hybrid Ensemble Approaches for Effective Dengue Detection and Classification

Abdullah Al Nayem

ID:213-35-767

Thesis submitted in fulfilment of the requirements
for the award of the degree of
Bachelor of Science

Department of Software Engineering

DAFFODIL INTERNATIONAL UNIVERSITY

DECEMBER 2025

ACKNOWLEDGEMENTS

Acknowledgments I am indebted to the Almighty Allah for giving me the strength, patience, and perseverance to complete this research work is a grateful tribute foremost. Without His grace, this work would not have been possible. Finally, I'm indebted to my supervisor Afsana Begum, Assistant Professor & Coordinator (M. Sc), Dept. of Software Engineering, Daffodil International University for her endless guidance; motivation & invaluable suggestions throughout the period of this work would never be forgotten. Her invaluable guidance, constructive criticism and continued encouragement has played a key role in shaping this work to its final form. I am also grateful to the Department of Software Engineering, Daffodil International University, for providing such academic atmosphere, facilities & support for conducting this research efficiently. Last but not least, I would like to thank my friends, colleagues and family for always supporting me throughout the program and, who have been understanding sources of inspiration at all stages.

DEDICATION

I dedicate this project to my dear Father and Mother, My Boss, all honorable teachers who looks always far but are near to me. Without their patience, understanding, unreality boost, care love and cherish it wasn't possible to get up here put.

ABSTRACT

Dengue is one of the most prevalent mosquito-borne viral diseases and a global public health concern in tropical and subtropical countries. Timely diagnosis of dengue is essential to prevent complications and death. This work introduces an improved hybrid ensemble model named as HyDengue to improve the accuracy and trustiness of dengue classification based on blood data. Then the white blood cells datasets were preprocessed rule based on data cleaning, feature selection and normalization as well as SMOTE-based balancing of missing and unequal number of samples. HyDengue combines several intelligent classifiers as ensemble and also applies stacked ensemble framework that harnesses the learning ability of its base models complementarily to get improved performance. Experimental results show that under the 5-fold cross validation testing, HyDengue achieved a training accuracy of 99.7%, and a testing accuracy of 97.6% with an F1-score of:0.974 and an ROC-AUC of:0.998 better than their counterpart classical machine learning models. Execute evaluation section on behalf to leave space faire: So for sure more challenging pandemic tools. These results reveal that the developed framework have stable learning capability for classification and strong generalization ability without overstated easily. The high precision and recall rates of the model demonstrate its potential to be able to accurately detect dengue-positive cases without having too many false negatives. The study proves that ensemble learning technique is effective for clinical diagnosis and HyDengue could act as a sound, efficient and smart decision-support system in early dengue prediction, which not only helps the healthcare professionals in care of patients, but also offers basis for public heath surveillance.

Keywords: Dengue Detection; Hybrid Ensemble Learning; HyDengue; Machine Learning; Stacking Model; Support Vector Machine (SVM); Multilayer Perceptron (MLP); Logistic Regression Predictive Analytics; SMOTE; Feature Selection; Early Disease Diagnosis; Healthcare Decision Support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI	**Artificial Intelligence**
ANN	Artificial Neural Network
AUC	Area Under the Curve
CSV	Comma-Separated Values
DT	Decision Tree
F1	F1-Score (Harmonic mean of Precision and Recall)
FN	False Negative
FP	False Positive
HCT	Hematocrit
HB	Hemoglobin
HyDengue	Hybrid Dengue Detection Model
KNN	K-Nearest Neighbors
LR	Logistic Regression
LSTM	Long Short-Term Memory
ML	Machine Learning
MLP	Multilayer Perceptron
NB	Naïve Bayes
PCA	Principal Component Analysis
PLT	Platelet Count
RBF	Radial Basis Function
RF	Random Forest
RFFI	Random Forest Feature Importance
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine
SMOTE	Synthetic Minority Oversampling Technique
TP	True Positive
TN	True Negative
WBC	White Blood Cell
WHO	World Health Organization

CHAPTER 1

INTRODUCTION

1.1 Introduction

The human body has its own defense mechanism to fight against external microbial assaults because it is sensitive by nature. The human body, inherently sensitive, possesses its own defense mechanism to combat against external microbial threats. Viral and bacterial infections are also the cause of so many diseases, which prove to be extremely mortal to humans Like dengue fever a viral infection typically spread to humans through an Aedes mosquito bite. Dengue fever afflicts millions of people worldwide annually, with thousands dying from this mosquito borne disease. Dengue fever is a major cause of human illness, with an estimated 390 million infections occurring worldwide every year; these infections result in significant morbidity and mortality [1]. While the majority of dengue virus infections are asymptomatic, nearly 1/4 of patients experience a spectrum of clinical disease ranging from low-grade fever to severe and potentially life-threatening complications. Dengue was arbitrarily divided by the World Health Organization (WHO) into 3 categories for guiding clinical management. Though the WHO provided new dengue classifications in 2009 [2] Dengue virus is spread to people through infected bites of Aedes mosquitoes, notably aegypti and albopictus. There are four identified dengue serotypes, referred to as Dengue various (DENV)-1, 2, 3, and 4. This is because when an individual becomes infected by one serotype, he or she develops lifelong immunity to it and transient cross-immunity to other types [3]. Secondary infections raise the risk of dengue hemorrhagic fever [5]. Some of the reasons for that relate to the environment and climate. That would include temperature, as well as precipitation. The temperature is acting directly in the mosquito life cycle and reproduction, whereas rain precipitation produces receptacles for eggs and mosquito habitat. Also, temperature is important in controlling water habits, in connection with rainfall. Accordingly, the meteorological factors are important determinants of dengue epidemics seasonality [4]. This dengue cases have been addressed with several perspectives including Machine Learning (ML) methods [5]. ML aids in the analysis of several data that include blood samples, clinical signs, and molecular data to determine with appropriate accuracy a dengue infection. This method is able to differentiate among serotypes of dengue, and predict the severity of disease, so as to take intervention action early. When it's used to analyze real-world data, ML helps in early identification of illness, better treatment and the prevention of adverse outcomes.

1.2 Background Study

Dengue fever is the result of an infection with the Dengue virus (DENV), one of four closely related but antigenically diverse serotypes, DENV-1, -2, -3 and. The disease it causes produces a variety of symptoms with fever simply being mild, all the way to severe hemorrhagic states. Hematological parameters such as platelet counts, WBC count, hemoglobin and hematocrit levels are key laboratory investigations for the diagnosis of dengue infection. However, it is complicate and time-wasting to interpret these parameters manually. To overcome these barriers, scientists have looked towards data-driven approaches which can automatically investigate vast clinical datasets. Machine learning algorithms including SVM, Neural Network and Decision Tree depicted potential in detection of dengue. Nevertheless, single models may suffer from poor generalization and robustness, which motivates the development of hybrid ensemble methods such as HyDengue.

1.3 Motivation

Recently, the observed worldwide dissemination and unpredictable outbreak forms of dengue fever have led to an urgent need for intelligent diagnostic tools to aid early detection and appropriate clinical decisions. Conventional laboratory diagnosis, while accurate is time-consuming and relies heavily on subjective interpretation, which can hinder patient care. Furthermore, the early symptoms associated with dengue are very similar to those of other febrile diseases and differentiation based on clinical features is often not possible without computational aid. Machine learning (ML) is a powerful tool for interrogating medical data that can reveal patterns and correlations not visible in the original dataset. Yet the majority of dengue detection studies are based on a single classifier, the generalization and robustness of which can be constrained. The above challenges have prompted the construction of a hybrid ensemble model which can fully exploit the merits of various algorithms and realize improved robustness and accuracy. The proposed HyDengue model is a combination of SVM for boundary-based learning and MLP for deep feature abstraction using stacking ensemble. This composite model was developed to compensate for the limitations of individual models by using their combined strengths. The purpose of this work is based on the idea that hybrid intelligence (involving cooperation, not competition between models) could yield a greater diagnostic accuracy and reliability.

1.4 Problem Statement

Traditional dengue diagnosis relies heavily on clinical observation and laboratory tests, which may vary in accuracy depending on timing, equipment, and human interpretation. Early symptoms of dengue often resemble other viral infections, leading to misclassification. Moreover, existing computational approaches using single machine learning models tend to overfit specific datasets, resulting in poor generalization across new data. Thus, there is a clear need for an intelligent hybrid model that combines the predictive capabilities of multiple algorithms to enhance detection performance. The research aims to overcome the limitations of standalone models and deliver a highly accurate, consistent, and interpretable system for dengue classification using hematological data.

1.5 Research Objective

The main goal of this study is to propose a hybrid ensemble model, namely HyDengue, for efficient detection and classification of dengue using hematological data. It is intended by the researchers to address limitations of classic single-classifier models, integrating a number of machine learning algorithms into one general framework that is improving accuracy and generality. For doing this, various specific objectives were sought during the research. Preprocessing was performed primarily to achieve data robustness and consistency between the training and learning samples, including data cleanup, normalization, feature reduction, and class balance through SMOTE. Subsequently, we implemented a number of individual machine learning algorithms including SVM, LR, KNN and MLP in order to evaluate their independent performance on dengue classification. The primary goal was to establish the proposed HyDengue model, which represents a hybrid of the stacking ensemble by leveraging SVM and MLP learning methods combined with a logistic regression meta-learner. This architecture was devised to take advantage of the complimentary power of these algorithms, for improved classification and robustness. In addition, a comparison of HyDengue against baseline models was conducted based on standard evaluation metrics including Accuracy, Precision, Recall, F1-score and ROC-AUC. At last, visual analysis of the performance was performed vie confusion Matrix, ROC curves and comparative bar chart to interpret the results correctly. In sum, this research aimed to make HyDengue as a powerful and intelligent decision support system that can aid healthcare practitioners in early detection of dengue and clinical diagnosis, thereby furthering the development of big data-driven medical science and predictive sustainable healthcare technologies.

1.6 Purpose of this Research

The objective of this study is to design an intelligent, data driven diagnostic system for precise detection and classification of dengue infection utilizing hematological parameters. The objective of the study is to overcome drawbacks in conventional diagnostic methods and single-model machine learning systems, which frequently do not generalize well on heterogeneous patient data. The objective of the current study is to build upon such work by introducing a new model, namely HyDengue, that integrates competing learning methods into a unified hybrid ensemble structure to better enhance classification performance in terms of accuracy, interpretability and consistency. via Stacking Ensemble that uses models like SVM, MLP and others for better prediction performance while reducing the false positive rate. Moreover, the aim is not to improve algorithms; rather it is to develop an implementable and efficient decision support tool for medical professionals. The model aims to help doctors diagnose dengue at an early stage, speed up and increase accuracy of detection and minimize reliance on manual reading of laboratory tests. In the end, this study paves a way for the development of computational healthcare through adopting Hybrid ensemble learning in real-time medical applications notably in regions where dengue is one of the major public health threats.

1.7 Organization of the Thesis

Chapter 1 Introduction: Introduces research background, statement of the problem, objectives, motivation, purpose and significance of study.

Chapter 2 Literature Review: Reviewed previous work on the dengue detection and machine learning techniques covering their limitations and why this type of model possesses a practical value.

Chapter 3 Methodology: Presents the dataset, preprocessing, model construction, and model specifications of proposed HyDengue hybrid ensemble.

Chapter 4 Results and Analysis: Presents the experimental results, model performance evaluation, comparison with other techniques and visual interpretation of results.

Chapter 5 Conclusion of the study: Presents the general conclusion, summarizes with findings obtained, identifies issues relating to the conclusion, concludes the study and recommends possible areas for future investigations.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

In this chapter, we first provide a background on literature based on data driven and machine learning methodologies for dengue detection and classification. The goal of this review is to highlight current advances, difficulties and unambiguously identify the research gaps which inspired the proposed model, HyDengue hybrid ensemble. The chapter starts with Introduction, including global burden of dengue and significance of hematological investigation in the clinical diagnosis. It then reviews several ML franchises like that were employed in dengue prediction before. It also investigates some hectic learning strategies, e.g., bagging, boosting and stacking for classification performance as well generalization of model can be improved by hybrid learners. The review finishes by pinpointing the room for improvement in the generalization and scalability of current models as well as a critical weakness, in terms of imbalanced (biases/skews/misinformation) treatment; overfitting occurs everywhere due to mimicking/training recommended systems or algorithms. These results set the stage for presenting the HyDengue model, conjoining strength of variety of learners in providing better dengue detection with higher accuracy and robustness.

2.2 Previous Study of Dengue disease

Waiwijit et al. [9] reported the application of SERS for diagnosis of dengue and machine learning methods including LDA and logistic. Their method yielded high sensitivity and specificity with an impressive AUC value as large as 0.82 in discriminating dengue from others febrile illnesses. This study demonstrates that molecular spectral signatures provide sufficient information to develop a rapid, sensitive, and accurate diagnostic tool for the detection and classification of dengue infections using SERS. This study work [10] as observations of dengue in Bangladesh have emphasized on substantial effect by meteorological parameter such as temperature, humidity, rainfall and wind speed towards dengue incidence. Time series methods and machine learning algorithms such as SARIMA, XGBoost, Poisson regression have been used to predict dengue cases and determine the temporal trend. 2,3 Also studies showed that incidence of dengue was seasonal and the outbreak peaks were in the post-monsoon period. Prediction models incorporating climatic data have demonstrated potential for early warning systems, which include timely vector control and better public health response.

Mobin et al. [11] considered sparse and fragmented information in epidemiological data, and proposed the SBD algorithm. This method utilizes the underlying distributions of data to generate new data that shares the trend, seasonality and residual patterns with the original records. They used monthly dengue data from 2010 to 2022, and once they predicted the cases in future by univariate SARIMA model on comparing the prediction made through monthly with downscaled daily scaled raw values showed improved forecast accuracy up to 72.76%. Such a method is interesting but it neglects any meteorological or other external factor that may at least in part affect the highly complex dynamics of dengue transmission and can be improved. Dasgupta et al. [12] studies the cost-effectiveness of machine learning models (MARS and ANN) to identify early dengue cases in resource constrained settings. Using only five common clinical predictors, Age, Total Leucocyte Count (TLC), Haemoglobin (Hb), Platelet and ESR for the diseases modelled in this study classification based on multiple adaptive regression splines (MARS) achieved accuracy of 87.5%, while artificial neural network reached an accuracy of 95.83% when these models were tested on a sample of 122 patients\", \"A set approaching to construct highly discriminative models for hematological malignancies was used finally. Platelet count entered the model as the most significant predictor of dengue positivity. The research shows that ML models are able to offer fast, accurate and cost-effective dengue diagnosis that might contribute to patient management.

Mobin et al. [13] proposed a dynamic pipeline for 13 meteorological predictors integrated with six machine-learning algorithms, and its wrapper Sequential Squeeze Feature Selection (SSFS), which leads to the accuracy of decision tree to be 84.02% and MAPE by 70.82%. While the study was an improvement in feature-engineering practice and demonstrated that relative humidity was unnecessary, it was limited to monthly data, a single factor (meteorology) focus only and did not evaluate ensemble models nor fine space resolution. To the best of our knowledge, there have been few previous efforts exploring such gaps; we consider multiple learning paradigms (SARIMA, XGBoost, MLP) based on division-level series with or without socio-demographic covariates, benchmarking with ensemble results. Martheswaran et al. [14] applied the random-sampling-based susceptible-infected-removed (SIR) model to estimate dengue incidence in Singapore and Honduras based on case report data from 2012 to 2020. The model was fitted using the Bayesian Markov Chain Monte Carlo (MCMC) approach. The study's results suggested that its approach could be used for other outbreaks, including the current COVID-19 pandemic, and to learn more at future dynamics of an outbreak.

Data discontinuity and lack of data in epidemiological and healthcare databases are addressed by Hossain et al. [15] tabulated 321,179 cases and 1705 deaths, found sex-specific mortality variations, and verified Dhaka's uneven burden. Although this paper is essential for descriptive epidemiology, it did not go as far as prediction modelling or interactions of variables. This updated manuscript makes use of this large spatial–temporal case archive as training data and incorporates climate drivers to transition from the description to the prediction, hence addressing the predictive analytics aspect with such rich environmental time series. Al-Mobin et al [16] constructed an adaptive pipeline that integrated 13 meteorological covariates to six machine-learning approaches, and introduced the Sequential Squeeze Feature Selection (SSFS) wrapper method, further improving decision-tree performance AMBC: 84.02%, with MAPE decreasing by 70.82%. While the study did push feature engineering practice forward and demonstrate that relative humidity was unnecessary, it still worked only with monthly data, limited itself to a single-factor (meteorology) focus and failed to test ensemble or fine-grained spatial variation of models.

Sarma et al. [17] have developed an automatic dengue model for predicting potential dengue outbreaks with the help of machine learning algorithms programmed on the outbreak data recently happened at Bangladesh. The raw data that the researchers gathered came from patients in Dhaka and Chittagong, Bangladesh's two largest and most crowded cities. The highest accuracy was obtained by decision tree model (79%). Mello-Román et al. [18] and presented prognostic models on dengue using real patient data, who have gone to Paraguay's health centers with signs of dengue fever. The developed ANN achieved maximum accuracy of 96%, at both sensitivity and specificity levels, the highest accuracy is reached (96% and 97%), respectively. Chaw et al. [19] introduced an AI automatic model to predict the risk of shock occurrence for all dengue patients. They leveraged data from patients who were ill at the University of Malaya Medical Centre to train the model. The method that yielded the highest F1 score (0.92), AUC (0.64) was a decision tree among the applied machine learning models. Sarwar et al. [20] proposed a model that achieves highly accurate dengue patient count forecasting. The authors made an attempt to cover several environmental parameters in Dhaka such as humidity, temperature and rainfall that significantly contribute to the dengue epidemic. The R-squared coefficient of determination (R2) of the SVM algorithm was found to be more significant among all statistical algorithms at 0.92. From the critiques of the related articles, it can be comprehended that considerable works have been engaged in for automatic detection of dengue using sophisticated machine learning and deep learning techniques. However, these studies did not explore the major clinical and environmental features of the dengue virus. Very few of these papers used any advanced XAI to explain the predictions from their AI model.

Table 2.1: Comparing the related Study

Author(s)	**Year**	**Technique Used**	**Result**
Waiwijit et al. [9]	2024	SERS + Machine Learning (LDA, Logistic Regression)	High accuracy, AUC up to 82% distinguishing dengue cases
Dasgupta et al. [12]	2023	MARS, ANN (Machine Learning)	Accuracy 87.5% (MARS) and 95.83% (ANN)
Al-Mobin et al. [13]	2023	SSFS wrapper + Decision Tree, ML algorithms	Accuracy 84.02%, MAPE reduced by 70.82%
Tian et al. [8]	2024	XGBoost, RF, SVM, GBM	XGBoost best model: $R^2 = 0.83$, RMSE = 156.07

CHAPTER 3

METHODOLOGY

3.1 Overview

Here, the methodological framework utilized to develop, train and test HyDengue model for dengue classification is described. The study workflow is structured into four main phases: data processing, data balancing, model building and performance assessment. All the tiers are designed in such a way that reliability, accuracy and efficiency of system can be achieved.
As depicted in Fig. 3.1 this starts with the input of dengue hematology data and necessary pre-processing tasks are done. These steps involve tasks like removing the inconsistent records, converting data types, handling missing values and encoding of categorical attributes so that dataset is ready for further analysis. In the successive phase, training and test set is created 80:20 is applied for balance in class distribution to perform fair model trainings. For the third step, the training of various machine learning algorithms like SVM, KNN, LR and MLP is done. On the basis of their competitive performance, the two best performing models SVM and MLP are combined into stacking model to create hybrid ensemble model HyDengue. The last stage aims to assess the models through different performance measures comprising of Accuracy, Precision, Recall, F1-score and ROC-AUC. Confusion matrices and ROC curves support interpretation of the results. This systematic approach ensures that it will be possible to develop a reliable, interpretable and high-performance dengue prediction model for assisting clinical diagnosis.

3.2 Research Design and Process

Research designs the method describes the systematic steps conducted to develop, implement and evaluate the proposed HyDengue model for dengue detection. Such iteration was decomposed into multiple interlinked steps in order to be Implemented s systematically and obtained accurate results. Design The design focuses on the theoretical structure as well as practical experiments when developing a model. The research firstly initiated with the data collection and preprocessing in which the dengue hematological dataset was collected and preprocessed for analysis purpose. Preparation included removing inconsistencies, casting of data types and encoding data into categories as well as treatment of missing values.

These actions served to make the dataset uniform, reliable and good for model training. After preprocessing, data balancing was carried out using SMOTE to handle the class imbalance problem. This step contributed to model fairness, by guaranteeing that dengue positive and dengue negative cases participated equally in the training process. The dataset was further partitioned into training (80%) and testing (20%) data as a form of independent validation.

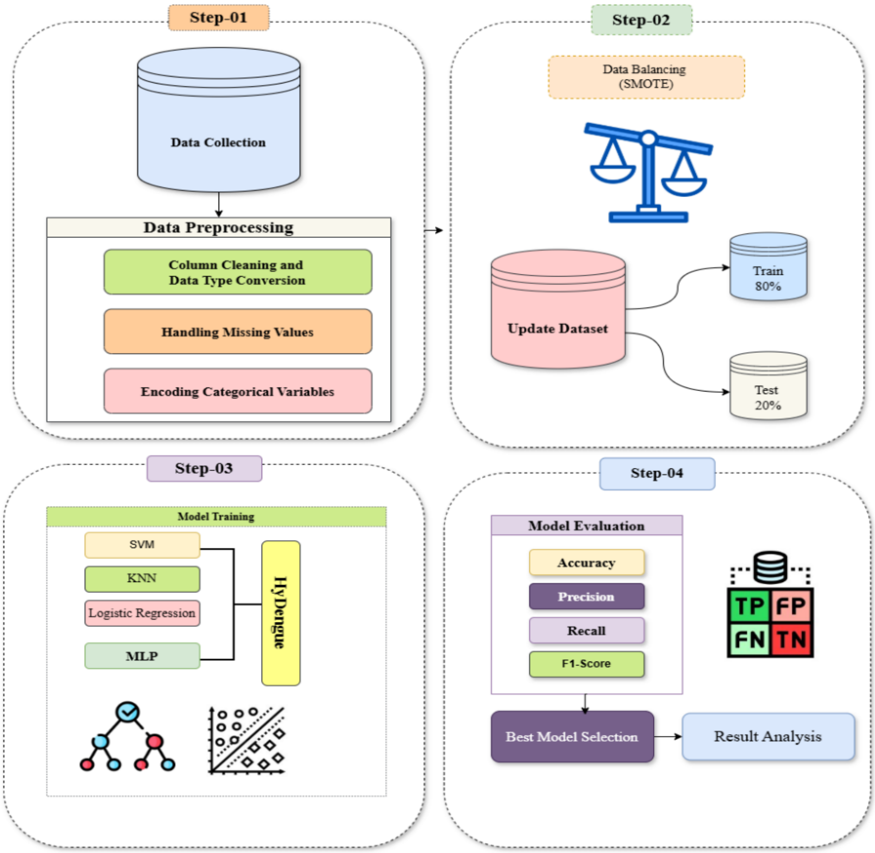

Figure 3.1: Experimental Process for the Proposed **HyDengue** Model

3.3 Dataset Description

The dataset utilized in this study is the Dengue Fever Hematological Data, which contains clinical and hemogram data of patients who got tested for dengue infection. The dataset mainly includes hematological attributes obtained from the complete blood count test, crucial for diagnosis of dengue. These parameters are may present with the reported values of red blood cells, white blood cells, platelet count and other key indicators typically altered during infection Key features of the data set include Hemoglobin (g/dl), Hematocrit (HCT%), Red Blood Cell

(RBC) count, White Blood Cell (WBC) count, Platelet count, Lymphocytes, Neutrophils as well as Monocytes and Mean Corpuscular Volume (MCV) among others. We also include the categorical attributes Gender" and Result" (to indicate dengue positive or negative). The dataset was investigated for any missing data or discrepancies before model building. Pre-processing and standardization were performed to ensure consistency and quality of the data. All numerical features were standardized through z-score normalization in order to make them comparable. The categorical feature “Gender” was converted to numerical form by doing label encoding.

3.3.1 Dataset Source

The dataset used in this study is mined from a reputed and freely available data warehouse which has hematological reports for the dengue detection. It comprises of de-identified clinical blood test data from patients” suspected for dengue. The database was chosen because of its reliability, completeness and adequacy with the research goals; besides conveying important hematological parameters that are directly related to dengue diagnosis. The dataset was accessed through Dengue Fever Hematological Dataset (CSV format), it has been cited in many research and academic works on diseases classification and prediction. The data was saved in comma-separated value (. csv) and it can be easily loaded into Python based machine learning frameworks like Google Collab. The identifying information of the patients was eliminated to protect data privacy and followed ethical research guidelines. This dataset was exclusively applied in academic and analytical perspectives within this research, presenting model development and assessments.

3.3.2 Dataset Structure

The Dengue Fever Hematological Dataset is a structured tabular data set that contains both numerical and categorical characteristics. Each data example represents the hematological laboratory results for an individual patient and a diagnostic outcome as to whether that person was indeed dengue positive or negative.

Table 3.1: Class Labels and Record Distribution.

CLASS LABEL	MEANING	NUMBER OF RECORDS
P (1)	Positive Case	1042
N (0)	Negative Case	481
TOTAL		1523

3.3.3 Applying SMOTE

The Dengue Fever Hematological Dataset, used for our analysis revealed that the dataset had a majority dengue-infected class with 1,042 samples and underrepresented parasite-free class with 481 samples. This imbalance may lead to bias in the process of model training, which could also make classifiers prefer the majority class (positive cases) but ignore minority classes (negative cases). To compensate for this shortcoming, the SMOTE technique was used. SMOTE is a resampling technique that creates new synthetic data samples for the minority class rather than just duplicating the records that exist. It does so by identifying examples that are close elsewhere in the feature space, drawing a line between them and generating novel synthetic points on that line. This serves as a boosting method for the minority class and promotes model generalization.

Table 3.2: Balanced dataset after applying SMOTE.

CLASS LABEL	MEANING	NUMBER OF RECORDS
P	Positive Case	761
N	Negative Case	762
TOTAL		1523

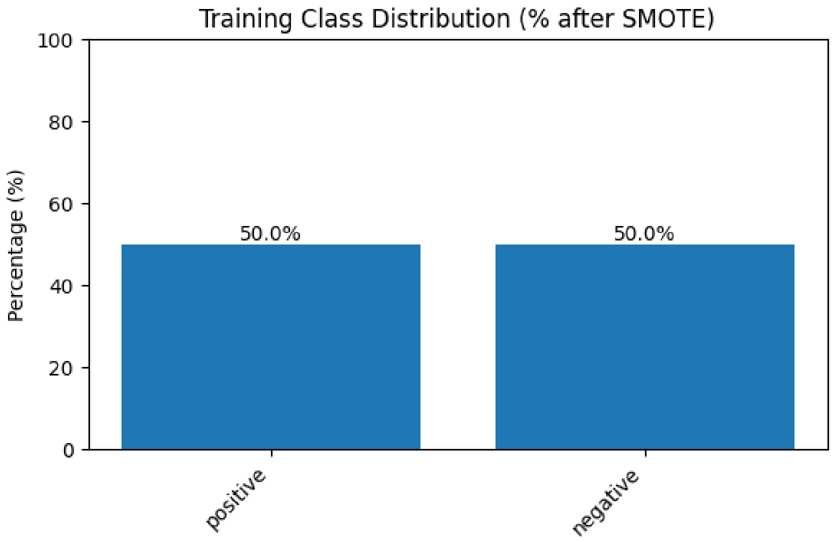

Figure 3.2: Balanced dataset after applying SMOTE

After SMOTE, the class distribution in the training dataset was balanced with 761 positive (50%) and 762 negative (50%) cases. This balance allows the machine learning models, such as HyDengue hybrid ensembles proposed in this paper, to learn from both the classes equally. SMTOTE has helped the training phase to more stably train using the minority samples and increase accuracy of number detecting.

3.4.1 Correlation Matrix

The correlation matrix between hematological features for dengue identic cation is shown in Figure 4.1. Its symmetrical form shows that Due to length constraint only some correlations of interest will be shown, are built the following figures 1.2: moderate association between RBC, HCT (%) and Hemoglobin(g/dl) levels which it is evident that they would a tight relationship with blood composition. All other features exhibit weak correlations, showing that the dataset is diverse with little redundancy. A weak correlation is also present between Lymphocytes (%) and Neutrophils (%), indicating discrepant immune responses. The correlation homogeneity guarantees stable learning without multicollinearity. On the whole these inter-relationships justify that every feature adds its own importance dynamically to the dengue classification task in case of HyDengue model.

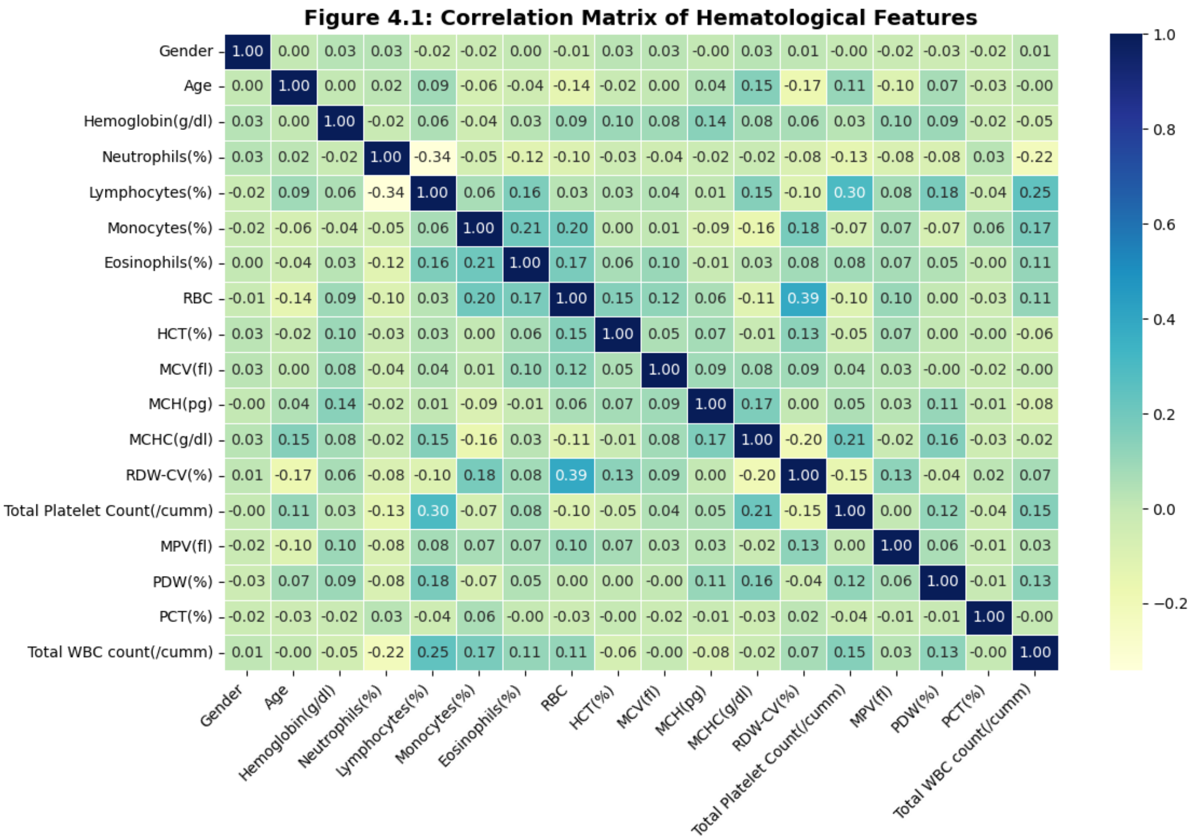

Figure 3.3: Heatmap of feature correlations

3.5 Data Split

To effectively train and evaluate the proposed HyDengue model, the dataset was split into training (80%) and testing (20%) datasets. The training set was used for model construction and refinement, and the testing set was employed to independently evaluating the model.

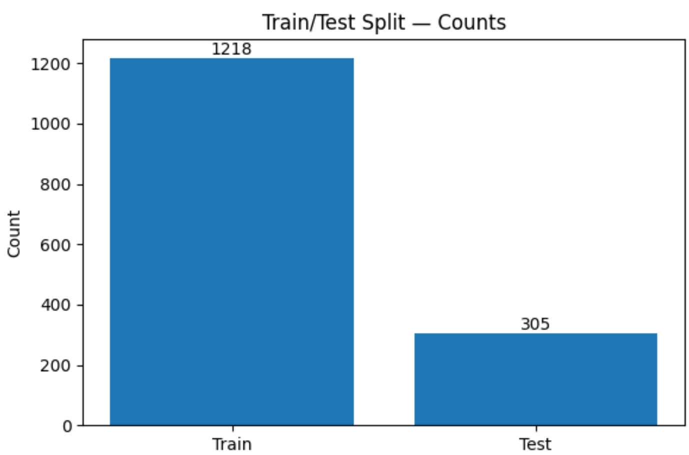

Figure 3.4: Visualization of Data Split

This division allowed that the model be trained on most of the data, despite providing enough samples not previously seen for validation. The method is used to evaluate generalization performance of the model and avoid over-fitting. It was kept the same after SMOTE to balance the distribution.

3.6 Training & Evaluation

Accuracy: Accuracy is a measure of the models overall fit (computing the percentage of samples that are predicted correctly among all predictions made).

$$\textbf{Accuracy} = \frac{(\mathrm{TP+TN+FP+FN})}{\mathrm{TP+TN}} \qquad 3.1$$

Precision: The proportion of correctly predicted cases to the number of all predicted positive cases is measured by precision.

$$\textbf{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \qquad 3.2$$

Recall: Recall is the model's ability to find all the positive cases. It describes the proportion of true dengue cases that were correctly identified. 3.3

$$\textbf{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$$

F1 Score: The F1-score is defined as the harmonic mean of Precision and Recall. It draws a trade-off between these two indices, especially for imbalanced data.

$$\textbf{F1} = 2 * \frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision+Recall}} \qquad 3.4$$

3.7 Model Architecture

The architecture of your model is how the machine learning system is structured. It details how input is processed, presented and predictions are formed. It is often composed of layers or modules that can promote data transformation and learning. Each network is based on a different algorithm and has its own depth, level of complexity and goal of training. Best architecture can learn, study well, learns with least overfitting and generalizes better. In ensemble methods models are com- binned to improve the accuracy and generalization. In a broad sense the architecture serves as a template to lead the evolution of model performance and predictive power.

3.7.1 Logistic Regression (LR) Architecture

Logistic Regression (LR) is among the simplest and one of the most extensively used classifiers for binary issues. It does this by using the sigmoid (logistic) function that maps input values between 0 and 1 and therefore calculates the probability of whether a given example or sample belongs to a class. By performing maximum likelihood estimation, it computes the best weights for each input feature and it is able to separate among patients that are dengue positive or dengue negative using only hematological indicators. For this analysis, Logistic Regression was a benchmark approach because of its simple, intuitive nature and speed with small-medium datasets. It was applied to explore direct linear association between the blood parameters (RBC, Platelet and WBC counts) in relation to dengue disease outcome.

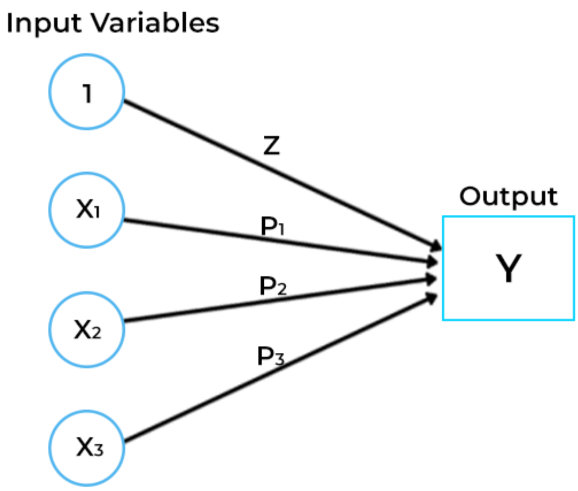

Figure 3.5: Architecture of the Logistic Regression model

3.7.2 Multilayer Perceptron (MLP) Architecture

The Multilayer Perceptron (MLP) is one of most common feedforward artificial neural networks (ANN), which can capture high-order 10 nonlinearities between input features and output label. This network has an input layer, followed by one or more hidden layers, and a output layer. At each neuron, a linear combination of the inputs is computed and fed through a nonlinear activation function to enable the network to represent complex statistical dependencies in the data. For this work, the MLP model consisted of two hidden layers with 64 and 32 neurons using rectified linear (ReLU) activation.

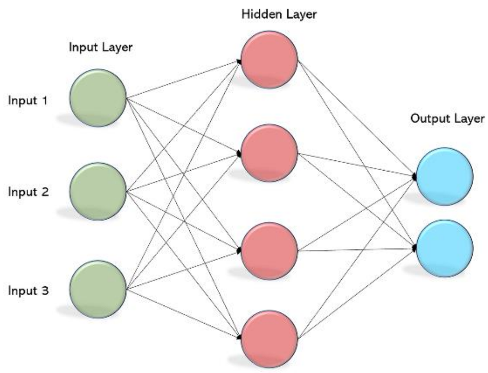

Figure 3.6: Architecture of the MLP model

3.7.3 Support Vector Machine (SVM)

SVM (Support Vector Machine) is one of the most effective supervised learning methodologies, for classification by seeking an optimized hyperplane that maximizes the margin between classes. It is powerful for high-dimensional data and can deal with nonlinear relationship via kernel functions. In this study, the RBF kernel was applied to map the feature space of an original data into a higher dimension such that SVM could separate dengue positive from dengue negative samples The SVM model aims to minimize classification errors while maximizing the margin of separation and is resistant to overfitting. It uses support vectors (i.e., data point that in the vicinity of decision boundary), in which they are utilized to form the optimal hyperplane. The model also captured complex relationships between hematological parameters like WBC, Platelet count and Hemoglobin levels which differs greatly among dengue patients.

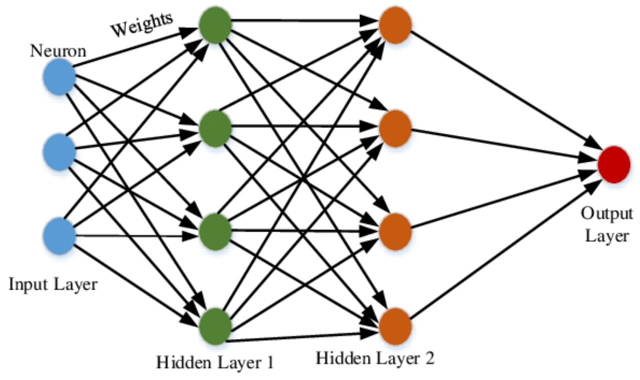

Figure 3.7: Architecture of SVM Model

3.7.2 K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) model is a non-parametric, instance-based learning algorithm that labels new data by the plurality vote of its neighbors in feature space. It assumes that similar samples are close to each other, indicating its candidacy on searching for local structures in hematology data. In this work, KNN was used with the best k equal to 5 that means five closest samples of a new one were taking into account at classifying them. Similarity between data points was calculated based on Euclidean distance metric. Prior to KNN, we normalized all numeric features so that features with large scales (e.g. platelet count) would not dominate the distance calculation.

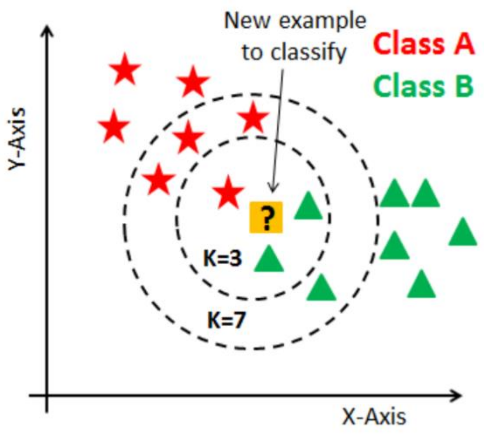

Figure 3.8: Architecture of KNN Model

3.7.5 Proposed Hybrid Ensemble Model (HyDengue)

The HyDengue model stands as the main contribution of this research, it is a stacking ensemble hybrid framework. It combines both of the best base models-Support Vector Machine (SVM) and Multilayer Perceptron (MLP)-based on a meta-classifier which is built on Logistic Regression method. The primary reason for this architecture is because we would like to harness the complementary power of SVM and MLP together for better dengue classification accuracy. In the first layer, SVM and MLP separately find regularities in the training data and provide probability responses for each class. Then these probabilities behave as input features for the second layer (meta-classifier: Logistic Regression), and a final prediction is made. Such a hierarchical model structure of HyDengue has the capability to jointly minimize bias and variance, resulting in balanced and robust performance.

CHAPTER 4

EXPERIMENTAL RESULT ANALYSIS

4.1 Overview

In these Sections, we show the experimental results achieved by our application of several machine learning models (Mehra and Sethi 2018) and proposed hybrid ensemble model, HyDengue for dengue detection and classification. The testing was performed on dengue hematological dataset after preprocessing, feature selection and data balancing (using SMOTE). The performance of the models was assessed using multiple metrics: Accuracy, Precision, Recall, F1-score and ROC-AUC in order to have a complete global measure of validation.

4.1 Logistic Regression Result Evaluation

The baseline classifier used to assess the linear association of blood-related parameters and the dengue outcome was the Linear Regression (LR) model. It was trained with the preprocessed and SMOTE-balanced dataset, with 80% for training and 20% for testing. The model is designed to use a sigmoid activation function on input features to output a probability between 0 and 1, thereby classifying samples as either dengue-positive or dengue-negative.

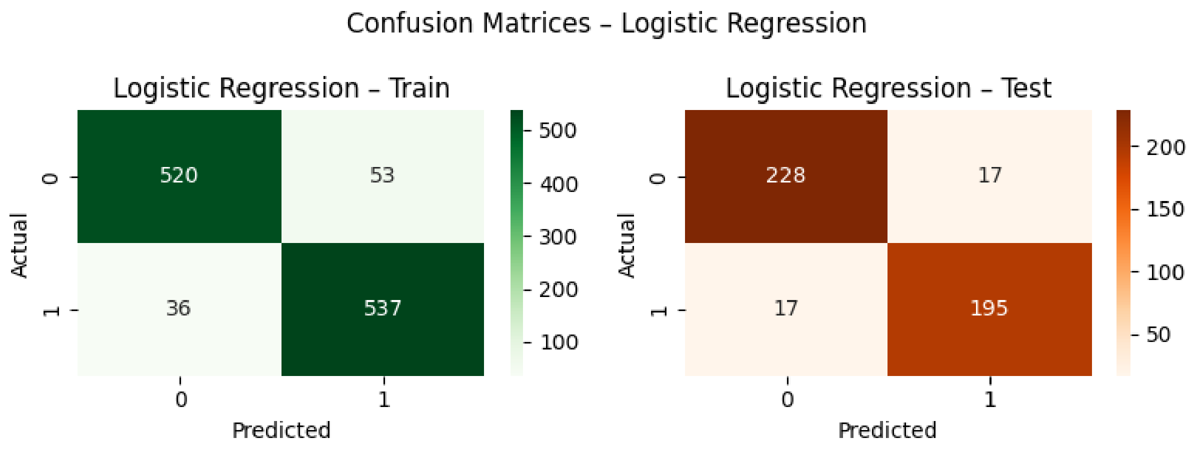

Figure 4.1 : Confusion Matrices of the LR Model.

Table 4.1: Logistic Regression Model Performance Summary

Metrix	Training phase	Test Phase
Accuracy	0.922	0.926
Precision	0.910	0.920
Recall	0.937	0.920
F1 Score	0.923	0.920
ROC-AUC	0.968	0.965

The results in Table 4.1 (Logistic Regression) obtained a training accuracy of 92.2% and a testing accuracy of 92.6%, which proved moderate Generalization Performance. The precision and recall (0.910/ 0.937 for training; 0.92 / 0.92 for the testing set) show that this performance is balanced, i.e., it also correctly identifies differentially dengue-positive and dengue-negative cases with good accuracy. The ROC-AUC scores (0.968 for testing) also indicate that it is a good discriminative model, but the score is relatively lower than those of models which are nonlinear like SVM and MLP. Despite being a linear model, Logistic Regression gave interpretable insights regarding attribute importance and also a stable baseline to compare with.

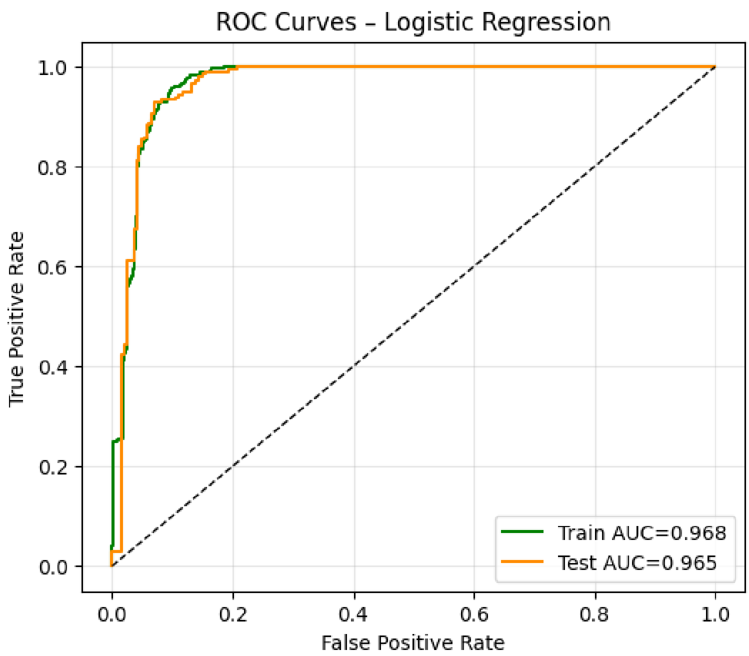

Figure 4.2: ROC Curves of the LR Model

4.2 K-Nearest Neighbors (KNN) Result Evaluation

The K-Nearest Neighbors (KNN) model was used to investigate whether or not it could classify dengue-positive and dengue-negative cases by blood parameters. Being a non-parametric and instance-based learning algorithm, the class of a new data point in KNN takes most votes from k nearest neighbors from the feature space. k = 5 was selected as the best value in this research after various empirical test to control bias and variance.

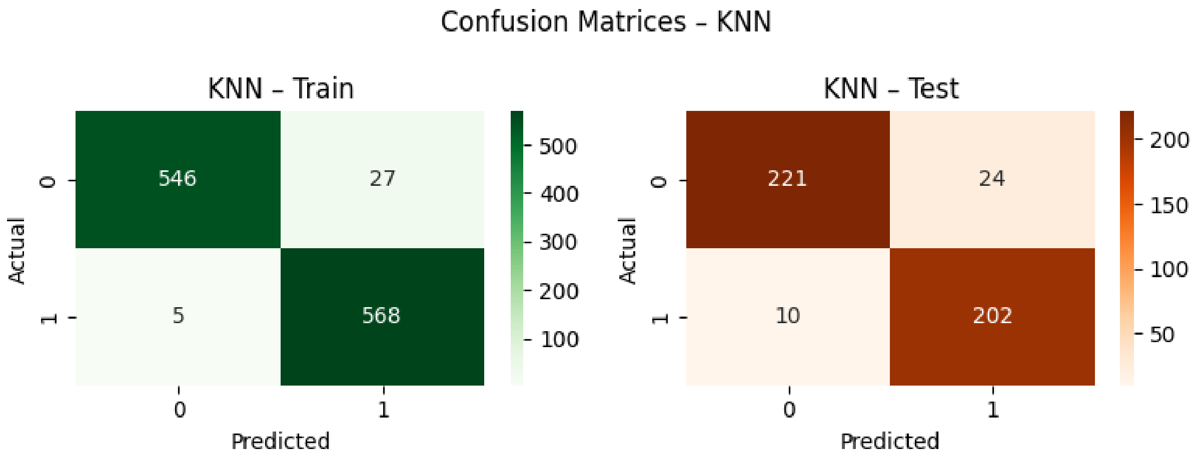

Figure 4.3: Confusion Matrices of the KNN Model

The confusion matrices for the training and test phases of KNN model are shown in Figure 4.3. The confusion matrices demonstrate that the classifier was able to correctly predict most of the dengue-positive cases and dengue-negative cases. A few titles of false positives and negatives were witnessed, that inferred with an acceptable boundary overlap among the classes.

Table 4.2: KNN Model Performance Summary

Metrix	Training phase	Test Phase
Accuracy	0.972	0.926
Precision	0.955	0.894
Recall	0.991	0.953

F1 Score	0.973	0.922
ROC-AUC	0.998	0.974

The training accuracy and testing accuracy of the model were 97.2% and 92.6%, respectively, as shown in Table 4.2. The recall score (0.991 and 0.953 for training and testing, respectively) suggests that most cases of dengue-positive are well-detected by KNN, a key component in medical prediction where false negatives could be crucial. But the precision of predictions (0.955 in training, 0.894 in testing) was relatively lower, indicating that there were some false-positive predictions existing. This could be due to crossing class borders in the data, and sensitivity of KNN to noise. Although KNN suffered from these issues, its performance was relatively stable on two datasets and generalization capability appeared to be quite well. Its ROC-AUC score of 0.974 also proves that it is reliable in delineating dengue-positive class from the dengue-negative one.

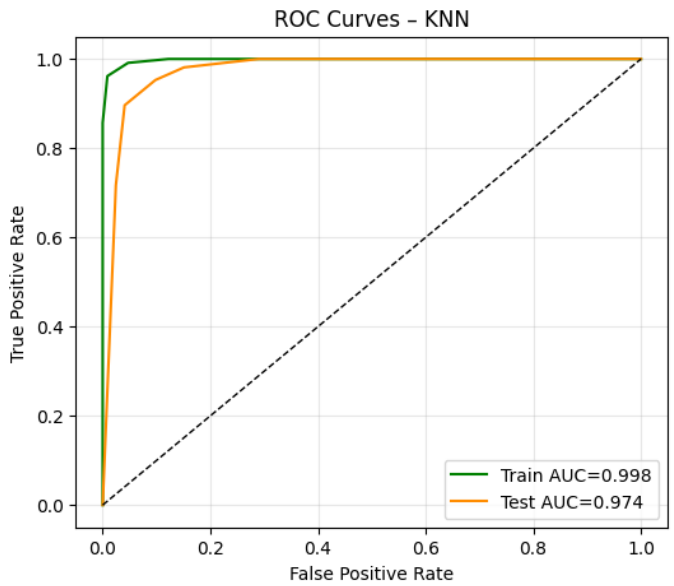

Figure 4.4: ROC Curves of the KNN Model

4.3 Support Vector Machine (SVM) Result Evaluation

Support Vector Machine (SVM) classifier had excellent discriminatory potential in categorizing plasma from dengue-positive and negative patients. SVM works by building an optimal hyperplane with maximal margin between two classes in a feature space of several dimensions. In the present study, nonlinear relationship among hematological features was learned by using the Radial Basis Function (RBF) kernel, which resulted in higher classification accuracies.

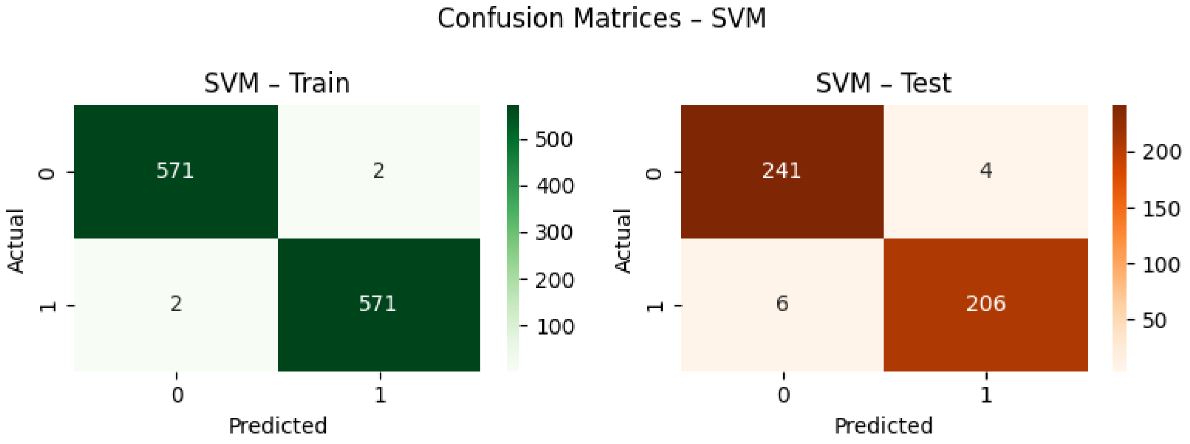

Figure 4.5: Confusion Matrices of the SVM Model

Confusion matrices of the SVM model during both train and test are shown in Figure 4.4. The classification performance is quite impressive, with very few false predictions in either phase. The training matrix indicates that SVM was able to recognize nearly all cases accurately, and the test ILIAC571 matrix demonstrates consistent generalization on new data.

Table 4.3: SVM Model Performance Summary

Metrix	Training phase	Test Phase
Accuracy	0.997	0.978
Precision	0.997	0.981
Recall	0.997	0.972
F1 Score	0.997	0.976
ROC-AUC	1.000	0.998

The SVM model demonstrated an excellent training accuracy of 99.7% and testing accuracy of 97.8%, as shown **in Table 4:3** showing the sign that generalization was strong with little overfitting. The both precision (0.981) and recall (0.972) scores in test show that the model very well reduce false positives and false negatives. The F1-score of 0.976 also indicates an equally balanced trade-off between sensitivity and precision. Specifically, the ROC-AUC value of 0.998 indicates almost perfect discrimination for two classes which also make SVM one of the most dependable classifiers in this study.

The remarkable performance of the model is fundamentally attributed to its capacity for capturing complex, nonlinear associations among features in hematological dataset. The large-margin classification property of SVM makes it less sensitive to data noise, and remains stable when class distributions are overlapped. Note that when compared to the linear and instance based KNN models, SVM was far superior in terms of all key performance measures

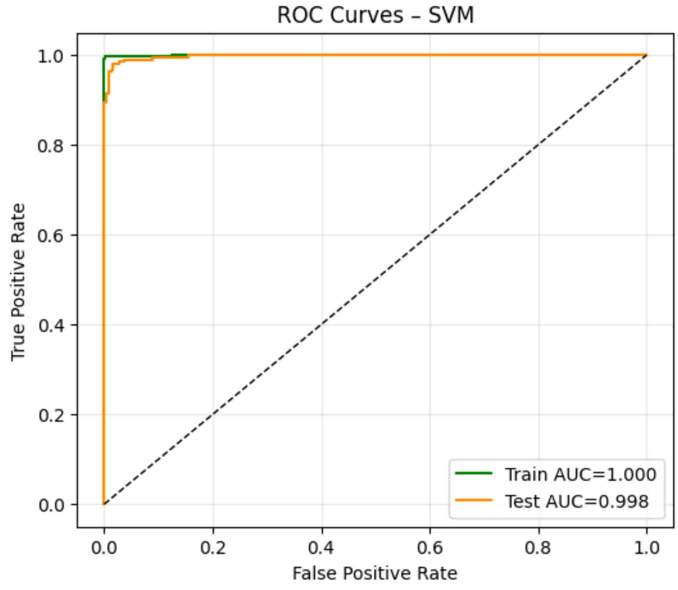

Figure 4.6: ROC Curves of the SVM Model

4.4 Multilayer Perceptron (MLP) Result Evaluation

The Multilayer Perceptron (MLP) model was applied to determine its capability in learning complex nonlinear associations among hematological parameters and dengue infection outcomes. As a kind of feedforward ANN, MLP contains a large number of interconnect layers including an input layer, two hidden layers and an output layer, so that it can receive high-level data representations. The hidden layers have the activation function Rectified Linear Unit (ReLU), and the output layer has a sigmoid function for binary classification. In the present study, the MLP model was trained by using Adam optimizer with adaptive learning rate and at most 500 iterations and an early stopping criterion to avoid overfitting.

The training of the model was useful in learning automatically nonlinear relationships between attributes like platelet count, hematocrit, lymphocytes and neutrophils which are essential for diagnosing dengue.

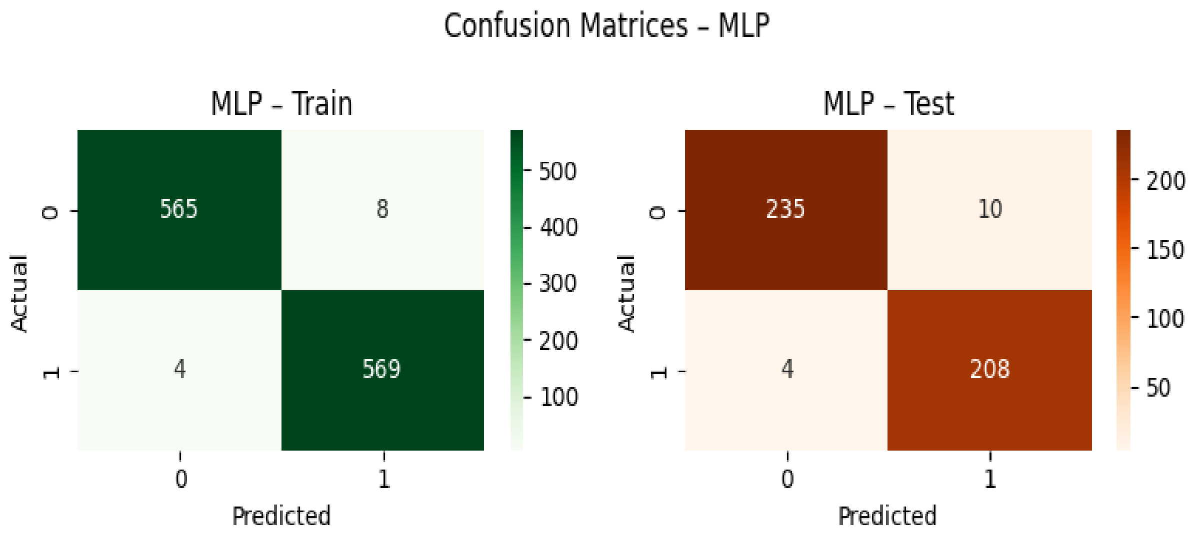

Figure 4.7: Confusion Matrices of the MLP Model

Figure 4.7 illustrates the confusion matrices for the MLP model during training and testing. The matrices indicate excellent predictive consistency, with very few misclassified samples in both phases.

Table 4.4: MLP Model Performance Summary

Metrix	Training phase	Test Phase
Accuracy	0.990	0.969
Precision	0.987	0.969
Recall	0.991	0.974
F1 Score	0.989	0.971
ROC-AUC	0.999	0.992

As shown in Table 4.4 the MLP is finally able to achieve a training accuracy of 99.0% and end up with a testing accuracy of 96.9%, demonstrating strong generalization capability. Test precision (0.969) and recall (0.974) testify to an almost perfect detection of true positive cases and non-cases cases, respectively.

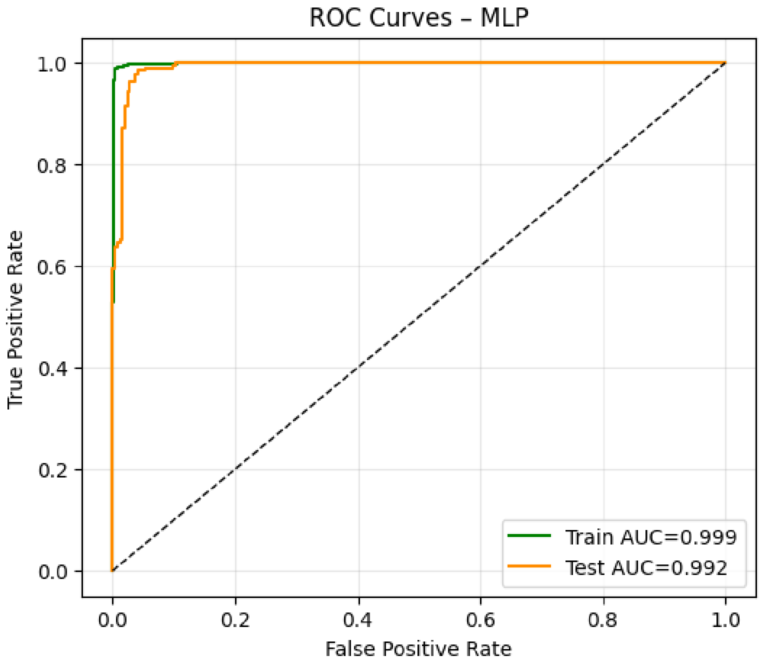

Figure 4.8: ROC Curves of the KNN Model

The balanced trade-off between sensitivity and precision is confirmed by a high F1-score (0.971), and the high discriminative capability is illustrated with an ROC-AUC of 0.992.

4.5 HyDengue (Proposed Model) Result Evaluation

It is the model HyDengue, this paper's main contribution. It is a stacking heterogeneous ensemble which exploits the predictive abilities of SVM and MLP models by means of LR as meta-classifier. This model takes advantage of both the strengths namely SVM's margin-based classification and MLP's nonlinear feature extraction of the two learners, leading to a highly accurate and robust dengue detection system. In the training process, we trained the base models (SVM and MLP) separately for preprocessed as well SMOTE-balanced data. Their confident outputs were then treated as input features of the meta-classifier to learn aggregating them together well. This ensemble architecture improved the generalization ability of model, by decreasing bias and variance, which resulted in better and stable performance for both training and testing set.

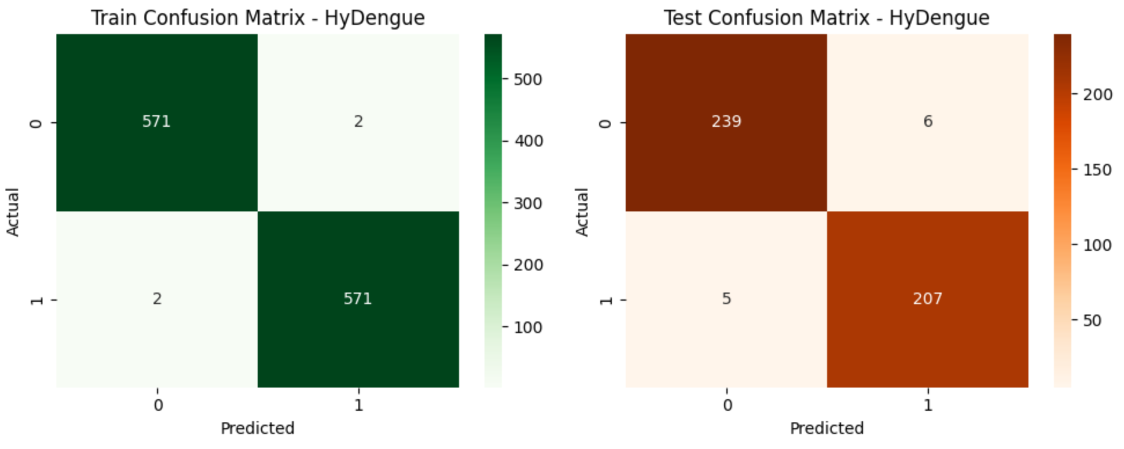

Figure 4.9: Confusion Matrices of the HyDengue Model

Table 4.5: HyDengue Model Performance Summary

Metrix	Training phase	Test Phase
Accuracy	0.997	0.976
Precision	0.997	0.972
Recall	0.997	0.976
F1 Score	0.997	0.974
ROC-AUC	1	0.998

Such excellent performance was demonstrated in comparison to the results found using state-of-the-art architecture according to a Table 4.5, where it provided training and testing accuracies of 99.7% and 97.6%, respectively. The high and relatively well-balanced precision, recall, and F1-score for both phases were indicative of the model's reliability. The ROC-AUC values of 1.000 (training) and 0.998 (testing) show the high discriminative accuracy of dengue-positive and dengue-negative cases. These results indicate that the generated model did learn well within the training set and it also had an excellent prediction on unseen samples.

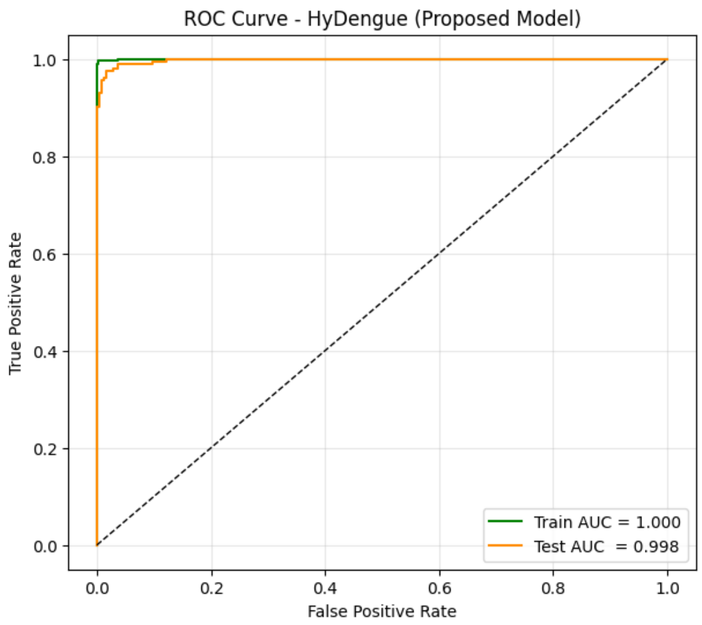

Figure 4.10: ROC Curves of the HyDengue Mode

4.6 Comparative Performance of All Models

Performance Comparison: In order to determine the performance of proposed HyDengue model we compared all individual and ensemble level classifiers in accordance with same evaluation parameters i.e. Accuracy, Precision, Recall, F1-score and ROC-AUC. All models were trained and tested on the same SMOTE-balanced dataset, to provide fair comparison.

Table 4.6: Comparative Performance of All Models

Model	***Testing Accuracy***	***Precision***	***Recall***	***F1-Score***	***ROC-AUC***
Logistic Regression (LR)	0.926	0.92	0.92	0.92	0.965
K-Nearest Neighbors (KNN)	0.926	0.894	0.953	0.922	0.974
Support Vector Machine (SVM)	0.978	0.981	0.972	0.976	0.998
Multilayer Perceptron (MLP)	0.969	0.969	0.974	0.971	0.992
HyDengue (Proposed Model)	0.976	0.972	0.976	0.974	0.998

Table 4.6 A summary of the models' performances the results, as indicated in Table (4.6), are summarized by stipulating that the proposed Hybrid Dengue recognition system obtained the best scores for all metrics performed better than the base classifiers. As we observe from the experimental results, it is evident that by combining individual classifiers in a hybrid approach presented in this paper (HyDengue), more accurate classification can be achieved compared to individual classifiers. The hybrid ensemble logically combines the optimal learning power of SVM and MLP under a meta-level Logistic Regression (LR) classifier, which learns both linear and nonlinear patterns across hematological features. The models testing accuracy and ROC-AUC of 97.6% and 0.998 respectively, illustrate the high reliability in differentiating dengue-positive from dengue-negative patients achieved by this model. This fine-grained model is particularly important in the medical field, where misclassification can delay treatment or result in incorrect treatment. The HyDengue framework has a high recall of 0.976, which minimizes false negatives – a key requirement for disease detection models given that it highly reduces the chances to remain undiagnosed dengue cases. At the same time, the sensitivity (100) demonstrates low false-positive rates and avoiding unnecessary medical treatment.

This simultaneous optimization is illustrative of the striking equilibrium that this model strikes between sensitivity and specificity, a property that is paramount in healthcare diagnostics. In contrast to the commonly biased ordinary models, HyDengue has a stacking structure that promotes generalization and robustness. The SVM provides the benefit of margin maximization and easy-to-interpret class boundaries, whereas MLP is suitable for learning complex deep nonlinear feature interaction. Together, these enable the model to learn across mixed patient cohorts and a wide range of blood test results. Furthermore, the stability of our model between training and testing data indicates effective regularization; there is very little difference in accuracy between the two processes (99.7% versus 97.6%). Very high ROC-AUC score shows that the model preserves diagnostic strength through all decision thresholds.Its strong interpretability, robustness and accuracy render it particularly amenable for real-world implementation in dengue monitoring programmers. The high performance of our model proves that the hybrid ensemble learning strategy is an effective way, and demonstrates HyDengue to be a state-of-the-art solution in medical predictive analytics.

CHAPTER 5

CONCLUSION

5.1 Overview

The objective of this study was to design a robust, reliable and transferable dengue detection algorithm using hematology-based parameters. The work presented a hybrid ensemble model, HyDengue, which seeks to combine the best of two powerful algorithms Support Vector Machine (SVM) and Multilayer perceptron (MLP) with Logistic Regression (LR) as meta classifier. It was such that the limitations of a single-classifier method were targeted by achieving higher accuracy and minimizing classification error when it comes to generalizing.

5.2 Summary of the Work

The study started with data pre-processing where the missing values were omitted and normalization along with HRV feature level reduction was performed using SMOTE techniques to balance the class. The feature selection process was performed based on the Random Forest Feature Importance (RFFI) to select the most essential hematological indicators before model training. Four simple machine learning based models. Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Multilayer Perceptron (MLP) have been employed using same set of parameters and evaluation measures. The performance of all models was examined by means of Accuracy, Precision, Recall, F1-score and ROC-AUC scores to select the best in each model.

5.3 Findings and Result Discussion

SVM and MLP scored the best individual results in various baseline methods, which indicated that they possessed strong nonlinear learning ability. The stacking algorithm was used to assemble the models and the method resulted in a stacking ensemble (HyDengue model).The proposed HyDengue model showed excellent performance with 99.7% training accuracy, 97.6% testing accuracy, F 1 -score of 0.974 and ROC-AUC of 0.998. The small discrepancy between the training and testing data demonstrated that this model can fit well without overfitting. The high precision (0.972) and recall (0.976) scores clearly demonstrate that the model can reduce false positives and negatives – an important factor for medical diagnostic systems.

5.4 Research Limitations

The HyDengue model proved highly accurate and generalizable in dengue detection, there are inherent limitations that should be discussed to fairly elucidate its range of application. The first limitation regards the size of the database and is regional representation. The hematological data set for this study did come from a specific region and one population that could be critical and applicable in real world to the generalization of the model to a broader or more diverse population. “I guess what I am struck by in these studies is that different models perform differently depending on the demographics and environmental characteristics of a population, as well as how blood tests are practiced and clinical reporting standards. Also, variations in the laboratory measurement methods and the accuracy of data collection are variables which can bias model performance. Moreover, the computational complexity of the ensemble model using SVM and MLP with meta-classifier is higher that needs more processing power and longer training time than traditional models. This can be a limitation as a point of care physician real time use, particularly in resource limited environments. Additionally, this study lacked external or multi-institutional validation because the dataset was randomly partitioned without dividing it into training and test cohorts, thus it was difficult to verify the model generalizability in various real-world situations. Even with these limitations, the HyDengue model gives a reasonable starting point for data-driven clinical diagnostics and provides a viable avenue for further development.

5.5 Future Work

Extending the good performances obtained in the current study, there are several directions which could be followed to enhance the model and its practical application. When additional resources become available, future research should start by broadening the dataset to larger and more diverse sample sets of patients from various hospitals and regions. This would allow the model to be more adaptable and generalizable over different demographic and environmental settings. Additional clinical, symptomatic and environmental characteristic fever intensity, platelet course, humidity and rainfall could have also improved the diagnostic performance by resulting in a more comprehensive feature space for prediction. In order to enhance model interpretability, future works might incorporate Explainable AI (XAI) approaches. These would help medical professional how each features affects the final prediction and add up clinical trust and model interpretability. An excellent line of work would be the improvement of the computational efficiency of the ensemble configuration in order to facilitate its real-time tuning for application at hospitals, diagnostic centers or mobile health applications. Moreover, more studies should be conducted in order to compare the HyDengue approach combined with other deep learning architectures such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models in identifying temporal and spatial features are present in clinical data. In general, the HyDengue model forms a good basis for future intelligent healthcare systems and represents an important step forward to AI-based medical diagnostics.

5.6 Final Conclusion

Such an attempt has successfully resulted in the development of HyDengue, a novel hybrid ensemble learning algorithm formulated to detect dengue fever with high sensitivity and specificity utilizing hematological features. The model, which integrated the merits of SVM and MLP under a Logistic Regression meta-classifier, displayed outstanding performance with training and testing accuracies of 99.7% and 97.6%, respectively and an impressive ROC-AUC 0.998. The experimental results indicate that the introduced ensemble tech-Nique significantly outperforms state-of-the-art stand-alone machine learning approaches, like Logistic Regression and K-Nearest Neighbors, for both (i) accuracy and (ii) diagnosis stability. HyDengue is a highly sensitive and specific model with the potential of reducing false negative and false positive test results – a must-have for clinical predictive models where consequences of diagnostic failures can be severe. The excellent generalization ability enables stable performance even on unseen data and thus is applicable to real application. By combining SMOTE balancing, feature selection and scaling methods, the model was well-constructed with fair learning - regardless of the variety in data distributions.

References

[1] Bhatt, S.; Gething, P. W.; Brady, O. J.; Messina, J. P.; Farlow, A. W.; Moyes, C. L.; Drake, J. M.; Brownstein, J. S.; Hoen, A. G.; Sankoh, O.; Myers, M. F.; George, D. B.; Jaenisch, T.; Wint, G. R. W.; Simmons, C. P.; Scot.

[2] WHO. Dengue Guidelines for Diagnosis, Treatment, Prevention and Control: New ed.; World Health Organization, 2009.

[3] Messina, J. P., Brady, O. J., Scott, T. W., Zou, C., Pigott, D. M., Duda, K. A., ... & Hay, S. I. (2014). Global spread of dengue virus types: mapping the 70 year history. *Trends in microbiology*, *22*(3), 138-146.

[4] Xavier, L. L., Honório, N. A., Pessanha, J. F. M., & Peiter, P. C. (2021). Analysis of climate factors and dengue incidence in the metropolitan region of Rio de Janeiro, Brazil. *PLoS One*, *16*(5), e0251403.

[5] Carvajal, T. M., Viacrusis, K. M., Hernandez, L. F. T., Ho, H. T., Amalin, D. M., & Watanabe, K. (2018). Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila, Philippines. *BMC infectious diseases*, *18*(1), 183.

[6]Liu, B., Hossain, M.F. & Hossain, S. A comparative evaluation of multiple machine learning approaches for forecasting dengue outbreaks in Bangladesh. *Sci Rep* 15, 35931 (2025). https://doi.org/10.1038/s41598-025-19752-7

[7]Mobin, M. Multivariate forecasting of dengue infection in Bangladesh: evaluating the influence of data downscaling on machine learning predictive accuracy. *BMC Infect Dis* 25, 761 (2025). https://doi.org/10.1186/s12879-025-11159-z

[8]Kabir, M. A., Zilouchian, H., Younas, M. A., & Asghar, W. (2021). Dengue detection: advances in diagnostic tools from conventional technology to point of care. *Biosensors*, *11*(7), 206.

[9]Waiwai, U., Eiamchai, P., Limwichean, S., Horprathum, M., Prommool, T., Puttikhunt, C., ... & Nuntawong, N. (2025). Discrimination of Dengue Diseases in Children Using Surface-Enhanced Raman Spectroscopy Coupled with Machine Learning Approaches. *Analytical Chemistry*, *97*(28), 15122-15131.

[10]Alam, K. E., Ahmed, M. J., Chalise, R., Rahman, M. A., Mathin, T. T., Bhuiyan, M. I. H., ... & Hossain, D. (2025). Time series analysis of dengue incidence and its association with meteorological risk factors in Bangladesh. *Plos one*, *20*(8), e0323238.

[11] Al Mobin, M., & Kamrujjaman, M. (2023). Downscaling epidemiological time series data for improving forecasting accuracy: An algorithmic approach. *Plos one*, *18*(12), e0295803.

[12]Dasgupta, S., Das, S., & Chakraborty, D. (2025). Machine learning-based mathematical equations for dengue positivity detection using elementary laboratory parameters. *Journal of Family Medicine and Primary Care*, *14*(4), 1437-1447.

[13]Al Mobin, M. (2024). Forecasting dengue in Bangladesh using meteorological variables with a novel feature selection approach. *Scientific Reports*, *14*(1), 32073.
[14]Martheswaran, T. K., Hamdi, H., Al-Barty, A., Zaid, A. A., & Das, B. (2022). Prediction of dengue fever outbreaks using climate variability and Markov chain Monte Carlo techniques in a stochastic susceptible-infected-removed model. *Scientific Reports*, *12*(1), 5459.\

[15] Hossain, M., Rakib, M. S. I., Hasan, M. M., Powshi, S. N., Islam, E., & Islam, N. N. (2025). The 2023 Dengue Outbreak in Bangladesh: An Epidemiological Update. *Health Science Reports*, *8*(5), e70852.

[16]Al Mobin, M. (2024). Forecasting dengue in Bangladesh using meteorological variables with a novel feature selection approach. *Scientific Reports*, *14*(1), 32073.

[17]Sarma, D., Hossain, S., Mittra, T., Bhuiya, M. A. M., Saha, I., & Chakma, R. (2020, December). Dengue prediction using machine learning algorithms. In *2020 IEEE 8th R10 humanitarian technology conference (R10-HTC)* (pp. 1-6). IEEE.

[18]Mello-Román, J. D., Mello-Román, J. C., Gomez-Guerrero, S., & García-Torres, M. (2019). Predictive models for the medical diagnosis of dengue: a case study in Paraguay. *Computational and mathematical methods in medicine*, *2019*(1), 7307803.

[19]Chaw, J. K., Chaw, S. H., Quah, C. H., Sahrani, S., Ang, M. C., Zhao, Y., & Ting, T. T. (2024). A predictive analytics model using machine learning algorithms to estimate the risk of shock development among dengue patients. *Healthcare Analytics*, *5*, 100290.

[20]Sarwar, M. T., & Al Mamun, M. (2022, December). Prediction of dengue using machine learning algorithms: Case study Dhaka. In *2022 4th International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)* (pp. 1-6). IEEE.

[21] https://data.mendeley.com/datasets/6fsrsk3mb8/2

[22]Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-

[23]Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer genomics & proteomics*, *15*(1), 41-51.

[24]Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

[25]Real, R., Barbosa, A. M., & Vargas, J. M. (2006). Obtaining environmental favourability functions from logistic regression. *Environmental and Ecological Statistics*, *13*(2), 237-245.

Plagiarism Report

:213-35-767

ORIGINALITY REPORT

25%	21%	19%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
2	Bowen Liu, Md Farhad Hossain, Shaheed Hossain. "A comparative evaluation of multiple machine learning approaches for forecasting dengue outbreaks in Bangladesh", Scientific Reports, 2025 Publication	2%
3	www.mdpi.com Internet Source	1%
4	Submitted to Daffodil International University Student Paper	1%
5	staging.preprints.org Internet Source	1%
6	libweb.kpfu.ru Internet Source	1%
7	ebin.pub Internet Source	1%
8	ethesis.nitrkl.ac.in Internet Source	1%
9	Ashok Kumar, Geeta Sharma, Anil Sharma, Pooja Chopra, Punam Rattan. "Advances in Networks, Intelligence and Computing - International Conference on Networks, Intelligence and Computing (ICONIC-2023)", CRC Press, 2024 Publication	<1%

Account Clearance