DIAGNOSIS OF DENGUE FEVER USING MACHINE LEARNING ALGORITHMS

By

AMONIKA THAKUR
ID: 213-16-596

FINAL YEAR DESIGN THESIS REPORT

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computing and Information System
Major: Artificial Intelligence in IoT

Under the Guidance of

Ms. Sonia Nasrin
Lecturer
Department of Computing and Information System
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh
October 21, 2025

APPROVAL

This Thesis titled **"Diagnosis of Dengue Fever Using Machine Learning Algorithms"**, submitted by **Amonika Thakur**, ID No: **213-16-596** to the Department of Computing and Information System, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computing and Information System and approved as to its style and contents. The presentation has been held on 21 October, 2025.

BOARD OF EXAMINERS

Md. Sarwar Hossain Mollah **Chairman**
Associate Professor and Head
Department of Computing and Information System
Faculty of Science & Information Technology
Daffodil International University

Md. Nasimul Kader **Internal Examiner**
Assistant Professor
Department of Computing and Information System
Faculty of Science & Information Technology
Daffodil International University

Md. Mehedi Hassan **Internal Examiner**
Lecturer (Senior Scale)
Department of Computing and Information System
Faculty of Science & Information Technology
Daffodil International University

Dr. Saifuddin Md. Tareeq **External Examiner**
Professor
Department of Computer Science and Engineering
University of Dhaka, Dhaka

DECLARATION

I hereby declare that; this thesis has been done by me under supervision of **Ms. Sonia Nasrin, Lecturer,** department of Computing and Information System (CIS) of Daffodil International University. I am also declaring that this thesis or any part of there has never been submitted anywhere else for the award of any educational degree like, B.Sc., M.Sc., Diploma or other qualifications.

Supervised By

Ms. Sonia Nasrin
Lecturer
Department of CIS
Daffodil International University

Submitted By

Amonika Thakur

Amonika Thakur
ID: 213-16-596
Department of CIS
Daffodil International University

ACKNOWLEDGEMENTS

The successfully completions of this research was possible only thanks to the sincere cooperation, guidance, and contributions of many people over the past two semesters. We are deeply gratefully to everyone to who has assisted us in one way or anothar.

First, we expressed our heartfelt thanks and gratefulness to the almighty for His divine blessing making here it a possible for us to complete the **Final Year Design Thesis (FYDT)** successfully. I express my deepest respect and grattitude to my supervisor, my teacher **Ms. Sonia Nasrin, Lecturer,** Department of Computing and Information System, Daffodil International University, Dhaka, Bangladesh. His deep me a knowledge and skillful guidance in the research on **Diagnosis of Dengue Fever Using Machine Learning Algorithms** played an a important role in advancing my research. It was only through his endless patience, wise guidance, constant enccouragement, careffuly supervision, consstructive criticism, valluable suggestions, and repeated reading and revision of my draft that this research work was possible.

I express my sinccere grattitude to the Honorable Head of the Department of Computing and Information Systems, Sir, for his kind cooperation in the successful completion of this thesis of mine. At the same time, I would like to expresse my grattitude to the other teacher's of the department and the staff for their help and support's.

Also, I would like to thenks to all my classmete's for thair supporrt's and discusions while completing my courseworks.

Finally, I would like to exprass my deepest respects and gratitude to my parents for their constant prayers, encouragements, patiences, blessings and emotionals supports. Without their always continued supports, this journey would not have been possibles.

ABSTRACT

The study offers crucial information to academics and medical experts, directing the choice of the best modeling algorithms for infectious illnesses. The whole outcomes of my "Diagnosis of Dengue Fever Using Machine Learning Algorithms" proposal may be seen here. This article goes into detail about how the idea was transformed into a thesis.The objective of this project was to create a system that can accurately diagnosis a dengue using machine learning algorithms , millions of people world wide are infected with dengue fever every year, another virus carried by mosquitoes. Dengue outbreaks are relentless,control strategies must be innovative and aggressive. In this context, Machine Learning (ML) techniques appear to be a promising avenue for increasing our ability to predict the occurrence and spread of dengue fever. This study examines how well different machine learning techniques predict dengue fever using a well compiled dataset of 1,037 items and 12 attribute. Some assembly algorithms were used in this work: XGBoost, Random Forest, AdaBoost and CatBoost & Decision Tree, Naive Bayes, K-Nearest Neighbors(KNN), Support Vector Machine (SVM) models were used. CatBoost outperforms other methods studied with an amazing accuracy of 96%. This accuracy is a testament to the algorithm's ability to learn complex relationships in multidemonsial data sets, making it an excellent candidate for dengue fever diagnosis. The present study emphasizes the early diagnosis of dengue infection using machine learning techniques applied to a huge clinical dataset comprising 1,037 patient records and several hematological and biochemical characteristics. The major goal of this study was to develop and test a reliable automated model for identifying patients as Dengue Positive or Dengue Negative based on blood test parameters such as Platelet Count, WBC, RBC, HCT, Lymphocyte (%), and Neutrophil (%). We testted and commpared eight suppervised machine learning algorithms: CatBoost, XGBoost, Random Forest, Decision Tree, AdaBoost, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Naive Bayes. Each moddel was trrained and then evalluated on preprocessed data in order to juddge the moddels on important mettrics such as accuracy, precision, recall, and the F1-score. The testting results shoowed that the higheest accuracy was by CatBoost at 96.15%, followed by XGBoost with 95.19%, and Random Forest at 94.23%, refllecting that the perfformance of enssemble bassed algorithms is much better in handdling compplicated and nonllinear data patterns. The Decision Tree had an accuracy of 88.94%, while simppler models like Naive Bayes, KNN, and SVM had lower accuracies in here.

Table of Contents

CHAPTER

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ABBREVIATIONS
ML - Machine Learning
DT - Decission Trees
RM - Random Forrest
XGB - Extreeme Gradent Bosting
CB - CatBoost
AB - AdaBoost
NB - Naives Bayes
SVM - Supports Vectors Machines
KNN - K-Nearests Neighbors
SHAP - SHapley Additive ExPlanations
WBC - White Blod Cells
RBC - Red Blod Cells
HCT - Hematocrits

CHAPTER 1

INTRODUCTION

1.1 Introduction

Dengue fever is a viral infection transmit by mosquitoes and it has continued to be an serious a health concern worldwide, infecting millions of people each year. Since dengue epidemics are unpredictable, innovative and aggressive strategies need to be developed for prevention or containment. Once the objectives were defined, the identification of specific tasks was conducted, like developing data collection procedures and choosing significant variables contributing to the differentiation of patients as dengue-positive and dengue-negative. In addition, protocols for data collection from various hospitals and diagnostic facilities were included in all the planning to ensure the stability and diversity of the dataset. The procedure are followed a strict ethical principles to maintain patient confidentiality and compliance with data protection laws. A detailes a project timeline was made, a showing the essential processes of data pretreatment and feature selection, model training, and evaluation. A resource this studying a consisted of are the identifying the computational tools, programming environments like Python and Jupyter Notebook, and libraries necessary to apply machine learning algorithms. Early involvement with medical specialists and epidemiologists was scheduled in order to test the data and validate the detection model developed to be in line with the diagnostic practices for found in the real world. In general, this systematic planning phase laid a very strong foundation for conducting a well-coordinated and scientifically sound research effort. It ensured that each phase, from the data collection to model validation, was the conducted in a very structured a manner with the objective of developing an accurate, efficient, and ethical dengue detection system that would classify patients as Dengue Positive or Negative.

1.2 Motivation

Dengue Fever have focused basically on clinical and statistical methods, which are often limited because these methods are restricted to traditional methodologies. Some research has looked into applying machine learning, focusing on specific algorithms such as the MLP Classifier Algorithm and Support Vector Machines. However, the comprehensive integration of diverse datasets into the prediction models-including demographic, blood-related, and environmental datasets-is lacking in the literature. The current study aims to bridge this gap by making a comprehensive estimate of Dengue Fever and using algorithmic learning methods to spot intricate patterns in a dataset that has been comprehensively collected. The project currently under way aims to add new dimensions to the existing collection of data in order to bring the field closer to more reliable and timely detections of Dengue Fever outbreaks.

Gupta, Gaurav, et al. [5] highlighted is the need to the differentiate Dengue and its subtypes with speed and accuracy in the early stages of the development of illness. This ability to predict dengue illness can save lives through timely diagnosis and treatment. Microarray and RNA-Seq data have been crucial in developing Dengue prediction models on various dengue datasets. Traditional methods such as the support vector machines and Bayesian findings on various of dengue datasets lack the semantic heaviness to evaluate sentiment through text at the paragraph or phrase level. The random forest classifier showed the highest mean score value with thus the study advises using machine learning for predicting dengue infection.

1.3 Rationale for the Study

Equipment used for conducting this research includes various advanced data analysis and machine learning approaches that work together to create and evaluate models. This study are following algorithms in machine learning: Random Forest, Support Vector Machine, Decision Tree, Naive Bayes, K-Nearest Neighbor, AdaBoost, XGBoost, and CatBoost. The algorithms mentioned have been a selected due to their proved efficiency in classification tests and the capability to handle big, complex nonlinear interactions in clinical data.In order to prepare an optimal dataset and improve the model's performance, several data preprocessing and preparation techniques were implement, including data cleaning, normalization, feature selection, and encoding.The research also drew on general criteria of the assessment such as accuracy, precision, recall, and here F1-Score for the objective analysis and comparison of model performance. All experiments and implementation of the models were conducted in Python-Jupyter Notebook, using key libraries such as Pandas, NumPy, Matplotlib, and SHAP for the interpretability analysis. This instrumentation system forms the basis in coming up with a reliable, fast, and interpretable machine learning model for detecting dengue. Here combinened dataset and this experience with artificial intelligence, the investigation advances automated diagnostic methods to assist the doctors in correctly and timely identifying dengue infections.

1.4 Research Question.

i. How does the age here affect the risk of develop dengue fever?
ii. What is the impacted of gender on the a incidence of dengue fever?
iii. Is there are a relationship between mosquito a numbers and incidence of dengue?
iv. Different platelet count and patterns exist that can be related to cases of dengue fever?
v. Do specific symptoms, such as headache or joint ache, imply a greater risk for dengue infection?

1.5 Expected Outcome

This research is expected to yield and a very accurate and reliable machine learning based model that can diagnose dengue infection using hematological and clinical data. The system was will aim to classify the patients as either dengue positive or dengue negative based on crucial features such as, the platelet count, WBC count, RBC, hematocrit (HCT), lymphocyte percentage, and neutrophil percentage. The model to be produced will do this using advanced techniques; hence, the accuracy, precision, recall, and F1-scores are expected to be very good. Furthermore, there is a project plans to provide clear data visualizations, feature importance insights, and model interpretation by utilizing SHAP analysis for improved transparency and credibility in decision-making. This is will, in the end, bring about the rapid, data-driven and cost effective dengue detection that shall assist health practitioners in early diagnosis and enhance public health management.

1.6 Report Layout

• Introduction.

The introduction here are clearly presents the context of the research, objectives, and importance, which thoroughly attracts the reader and provides the framework for further investigation.

• Background.

An effectiveness introduction will a set the context of the research, its purpose, and its importance, thereby engaging the reader and establishing the framework for further investigation.

• Data Collection and Pre processing

The research entails a stringent data collection procedures to ensure factual and relevant information. These include demographic information such as Age and Sex, besides key clinical markers such as Platelet count, Neutrophil count.

• Research Methodology.

Here the main objective of the study is to come up the with a detection algorithm that will be able to classify such as dengue positive and dengue negative cases using hematological and clinical data.

• Experimental Results and Discussion

The technological system comprises a strong system with the ability to execute machine learning algorithms successfully.The development, training, and evaluation of models are made easier by the use be for develop a model sequentially of computer languages.

• Impacts on society's and the enviroonments

The ethiccal deployment of machine learning models involves the of unforeseen results, including patient misclassification or overdependencce on decisions. The approach thereby suggested has the aim of supporting.

• Summarry of conclusions, futuare researchs

Study emphasizes the early diagnosis of dengue infection using machine learning techniques applied to a huge clinical dataset comprising 1,037 patient records and several hematological and biochemical characteristics.

• References:

Major goal of this study was to develop and test a reliable automated model for identifying patients as Dengue Positive or Dengue Negative based on blood test parameters such as Platelet Count, WBC, RBC, HCT, Lymphocyte (%), and Neutrophil (%).

CHAPTER 2

BACKGROUND STUDY

2.1 Preliminaries

Dengue Fever projection would require an orderly collection of tools and materials to set up the experiment. The technological system comprises a strong system with the ability to execute machine learning algorithms successfully.The development, training, and evaluation of models are made easier by the use be for develop a model sequentially of computer languages like Python and with the use of libraries such as pandas, and matplotlib. The experimental results and subsequent analysis will lead to a better understanding of the strengths and weaknesses of each model and thereby provide an informed platform for optimizing the system and selecting of the propertied of a model on the basis of requirements and concerns. Once the objectives were defined, the identification of specific tasks was conducted, like developing data collection procedures and choosing significant variables contributing to the differentiation of patients as dengue-positive and dengue-negative. In addition, protocols for data collection from various hospitals and diagnostic facilities were included in all the planning to ensure the stability and diversity of the dataset. The procedure are followed a strict ethical principles to maintain patient confidentiality and compliance with data protection laws. A detailes a project timeline was made, a showing the essential processes of data pretreatment and feature selection, model training, and evaluation. A resource planning a consisted of are the identifying the computational tools, programming environments like Python and Jupyter Notebook, and libraries necessary to apply machine learning algorithms. Early involvement with medical specialists and epidemiologists was scheduled in order to test the data and validate the detection model developed to be in line with the diagnostic practices for found in the real world. In general, this systematic planning phase laid a very strong foundation for conducting a well-coordinated and scientifically sound research effort. It ensured that each phase, from the data collection to model validation, was the conducted in a very structured a manner with the objective of developing an accurate, efficient, and ethical dengue detection system that would classify patients as Dengue Positive or Negative.

2.2 Related Works

Most of the earlier studies related to dengue fever detection have focused on clinical and statistical methodologies, which most of the time suffer from dependency on conventional procedures regarding various dengue datasets. Machine learning has been the subject of some study, with focus being placed on specific algorithms like the Support Vector Machines and MLP Classifier Algorithm. The procedure are followed a strict ethical principles to maintain patient confidentiality and compliance with data protection laws. A detailes a project timeline was made, a showing the essential processes of data pretreatment and feature selection, model training, and evaluation. A resource planning a consisted of are the identifying the computational tools, programming environments like Python and Jupyter Notebook, and libraries necessary to apply machine learning algorithms. Early involvement with medical specialists and epidemiologists was scheduled in order to test the data and validate the detection model developed. Previous studies on the prediction of Dengue Fever have focused basically on clinical and statistical methods, which are often limited because these methods are restricted to traditional methodologies. Some research has looked into applying machine learning, focusing on specific algorithms such as the MLP Classifier Algorithm and Support Vector Machines. However, the comprehensive integration of diverse datasets into the prediction models-including demographic, blood-related, and environmental datasets-is lacking in the literature. The current study aims to bridge this gap by making a comprehensive estimate of Dengue Fever and using algorithmic learning methods to spot intricate patterns in a dataset that has been comprehensively collected. The project currently under way aims to add new dimensions to the existing collection of data in order to bring the field closer to more reliable and timely detections of Dengue Fever outbreaks.

Gupta, Gaurav, et al. [5] highlighted is the need to the differentiate Dengue and its subtypes with speed and accuracy in the early stages of the development of illness. This ability to predict dengue illness can save lives through timely diagnosis and treatment. Microarray and RNA-Seq data have been crucial in developing Dengue prediction models on various dengue datasets. Traditional methods such as the support vector machines and Bayesian findings on various of dengue datasets lack the semantic heaviness to evaluate sentiment through text at the paragraph or phrase level. The random forest classifier showed the highest mean score value with 8.72, and thus the study advises using machine learning for predicting dengue infection.

With a dataset ranging from 2010 to 2016, including demographic, regional, and meteorological various dengue dataset parameters, Majeed, Mokhalad A., et al. [6] focused on the use of machine learning to forecast Dengue fever cases in Malaysia. Six different LSTM models, with various dengue dataset, were developed and evaluated to forecast Dengue, with an average RMSE of 3.17 across different lookback periods. The SSA-LSTM model outperformed three traditional various dengue dataset models-SVM, DT, and ANN-in terms of RMSE and proved to be effective across most Malaysian states. The results indicate that, compared to other various dengue dataset models, the SSA-LSTM model forecasts dengue cases in Malaysia with the least average error.

Zargari Marandi, Ramtin, et al. [7] showed a machine-learning algorithm intended for various dengue datasets to detect signs of plasma leakage early in the course of the disease. Importantly, this paper keeps the real-world use of models on various dengue datasets with resource-constrained settings by making decision support on various dengue datasets for low-to-middle-income countries like Sri Lanka. The five easily accessible factors that can help decide on admitting probable cases of dengue into the hospital are identified through the model. The study has focused on model interpretability with various multi-metric assessments, Shapley additive explanations, and discusses fairness concerns. The final model attains an AUC of 0.80, PPV of 76.9%, NPV of 72.5%, specificity of 87.9%, and sensitivity of 54.8% on the test dataset, having important implications for serious consequences in patients having dengue.

Panja et al. [8] proposed the model XEWNet, which is a combination of a wavelet neural network with exogenous parameters that can reliably forecast dengue outbreaks in three diverse regions. The model is flexible and can incorporate exogenous climatic factors, as demonstrated by statistical tests of causality. Wavelet transformation is integrated into a neural network structure in the XEWNet model, which generates more reliable longer-term projections. It represents complex nonlinear associations between dengue prevalence and rainfall in a simple, computationally efficient, and mathematically interpretable way. Comparative studies have established that the proposed XEWNet is better than both the statistical, machine learning, and deep learning techniques in 75% of cases for both short-term and long-term estimates of dengue incidence.

Yang, Hsiu, et al. [9] applied all the methods from the explainable artificial intelligence approach called SHAP and Extreme Gradient Boosting. This study was aimed at evaluate on the spatial distribution of dengue case residences in Kaohsiung City, Taiwan, from the year 2014 to 2015. The main findings were that the proportion of residential areas and general roads were highly linked with dengue case residences nonlinearly. On the other hand, agricultural features were inversely correlated with dengue incidence. In an effort to prevent dengue, accurate prediction maps based on landscapes were developed and could be used to inform resource distributions and adjustment of control strategies.

Indhumathi, K. et al. [10] used electronic patient health data to predict and estimate the spread of contagious diseases and identify which ones would most likely spread rapidly. For this, the three main goals of the project are identifying deadly symptoms based on patient age, predicting disease development before it occurs, and forecasting sickness incidence using machine learning and deep learning algorithms. In the proposed model, the features such as selection and classification are done in a hybrid manner. While Random Forest with XG-Boost is used to classify diseases, Antlion Optimization Algorithm has been utilized for feature selection. It also examines the forecast of diseases on a daily and year-by-year basis, making estimations from past epidemic occurrences and climatic factors such as rainfall and temperature.

Rustom Al Nasar, Mohammad, et al. [12] utilized machine learning techniques, specifically Artificial Neural Networks (ANN) and Support Vector Machines (SVM), to forecast dengue illness. The figures were collected from a government hospital in Lahore, Pakistan, and contain records for diagnosing dengue illness (positive or negative). 70% of the data is allocated for training, while 30% is reserved for testing. On the other hand, agricultural features were inversely correlated with dengue incidence. In an effort to prevent dengue, accurate prediction maps based on landscapes were developed and could be used to inform resource distributions. Remarkably, the proposed model boasts an incredible 96.19% accuracy rate, outperforming earlier research attempts.

Sarma, Dhiman, et al. [13] discussed the critical need for the exact forecasting of dengue epidemics due to the unavailability of vaccines and antiviral drugs. This paper proposes a new machine learning approach for forecasting dengue fever using patient-based data, including symptoms, past medical histories, and diagnostic reports. The dataset is constructed from real-time data samples of Bangladeshi dengue patients. 70% of the data is used for training, while the remaining 70% is used for testing. The machine learning technique used for the proposed categorization model is DT and RF. In this model, the decision tree ultimately performs better than the random forest, with a standard accuracy of 79%.

Salim and Nurul Azam Mohd, et al. [14] investigated was there are best machine learning model for dengue mostly outbreak prediction in five districts with the highest incidence of dengue fever in Selangor, Malaysia, between 2013 and 2017. Each model takes into account climatic variables such as rainfall, humidity, wind speed, and temperature. The Support Vector Machine (SVM) is the better a among these, with 70% accuracy, 95% specificity, and 56% the precision. The sensitive of the test dataset in the SVM notably here goes up to 63.54%, an impressive raise from 14.4%, showing that initial data was unbalanced. Machine learning has demonstrated potential in predicting dengue outbreaks, and the variable relating to the week of the year figured as one of the most important predictors in the SVM model. Rahman et al., [18] estimated the number of female adult Aedes aegypti mosquitoes in northern Thailand through mapping the socio-economic, climatic change, dengue understanding, mindset, and habits (KAP), and landscape factors using machine learning. We evaluated a total of five models using supervised learning; of these, the random forest model showed the best performance in predictions based on different parameters. Using a dataset on dengue illness, the approach of Sanjudevi, D., et al. [19] focused on the application of classification techniques to enhance the accuracy of the prediction.

The proposed technique is compared with the decision tree using WEKA as a data mining tool. Four processes are involved in the study, which includes gathering data from the UCI repository, applying SVM and decision tree classification algorithms, selecting features by applying regression approaches, and evaluation of results using sensitivity (SE), specificity (SP), and AUC. The study recommends the use of advanced technology to carry out a more detailed analysis of dengue's effects. Damien K. Ming et al. [21] A study carried out in Ho Chi Minh City in Vietnam presented the results of supervised machine learning predictions from integrated data of adult and infant dengue hospital patients using ml models. The dataset was then divided into trainee and hold-out sets, after which ten-fold cross-validation was applied for hyperparameter tuning. The models were then evaluated on the hold-out set. From the final dataset, 5.4% of the total 4,131 patients were suffering from dengue shock syndrome, comprising 477 adults and 3,654 children. All the following were considered as predictors: hematocrit, weight, age, sex, the day the patient was admitted to the hospital, and the platelet parameters over the first fourteen days.

2.3 Comparison of leading solutions

Table 2.1: Accuracy Comparison between Papers

Author	Dataset Collection	Algorithm	Accuracy Result
Rammtin, Marandi Zargari,et al. [7]	Real	Machines Learnings	87.9%
Russtom Mohhammad, Al Nasars et al. [12]	Real	Machines Learnings	96.09%
Dhimman,Sarrmas, et al. [13]	Real	Machines Learnings	79%
Naiyars,Iqqbal, et al. [16]	online	Machines Learnings	92%

N A Mohde, Sallim, ,et al. [14]	Real	Machines Learnings	70%
My Papers	Real	Machines Learnings	96.15%

Table 2.1 gave a quick summary of research related to dengue. I observed that most of the studies used secondary data that is available to everybody. Therefore, when both the source of data and the implementation region are different geographical areas, there is no reason why a good model cannot achieve reasonable accuracy. Moreover, most of the datasets had not included enough number of features, whereas we incorporated to relevant fever specific features into our study by using the patient's personal, diagnostic, and symptom based information.

2.4 Scopes of the Problems

Dengue Fever remains on a big and there are complex problem with far and reaching repercussions. Dengue fever causes a high burden on the public health systems globally, with millions of people affected by the epidemic annually. Geographical spread of the disease were further and complicates the control, which demands comprehensive strategies for prevention and control.

Dengue Fever also has a high economic burden, especially for resource-poor settings. Global connectivity is accelerating the chances of spread of Dengue, which demands flexible prediction systems under constantly changing conditions. The scale increases because the worldwide warming will eventually affect vector behavior and disease transmission; hence, flexible and adaptive solutions are needed.

2.5 Challenges

The present study emphasizes the early diagnosis of dengue infection using machine learning techniques applied to a huge clinical dataset comprising 1,037 patient records and several hematological and biochemical characteristics. The major goal of this study was to develop and test a reliable automated model for identifying patients. The dynamic spread of Dengue Fever further complicates the model as it depends both on environmental reasons and human activities, making the model adapt to changeable circumstances. Besides, a lack of sufficient substantial and high-quality Dengue Fever forecasted datasets may restrict the model on the training and universality. In addition, accessing and using the real-world dataset is impeded by ethical concerns, specifically issues of data protection and informed consent. The complexity and clarity should be harmonious; the too-complex models would impede understanding and acceptability by the clinicians and the general public. These challenges mentioned above need to be considered during the successful development and implementation of a dependable Dengue Fever prediction model.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Indtroduction

This chapter summarizes the details of the research technique used in this study to determine dengue using machine learning algorithms. The methodology contains details about the step-by-step process that was used to ensure accuracy, consistency, and reliability to meet the set objectives of this study. Here the main objective of the study is to come up the with a detection algorithm that will be able to classify such as dengue positive and dengue negative cases using hematological and clinical data. In turn here several consecutive steps were are performed in the following order: data collection, data preprocessing, EDA, feature selection, model creation, evaluation, and interpretation. And each of these stages significantly has contributes to raw data into insights and forecasts. The pretreatment procedure ensures that the dataset is clean, consistent, and ready for analysis. Exploratory Data Analysis helps to the understand pattern, trend, and the variable interrelations, which is vital for good model creation. The chosen machine learning algorithms of Random Forest, XGBoost, CatBoost, and Logistic Regression are trained and tested to establish the accuracy of dengue case detection.

Finally here findings, datasets, and model results, which contributes to global scientific efforts toward the eradication of dengue. Here besides, the inittiative will observe ethical data mannagement and transsparency standards in order to retain public conffidence and accountability. Overall, such an approach to susttainability enssures that the dengue detection moddel continues to impprove into a sccalable.

3.2 Research Subjects and Instrumentation

The major goal of this study was to develop and test a reliable automated model for identifying patients as Dengue Positive or Dengue Negative based on blood test parameters such as Platelet

Count, WBC, RBC, HCT, Lymphocyte (%), and Neutrophil (%). We testted and commpared eight suppervised machine learning algorithms: CatBoost, XGBoost, Random Forest, Decision Tree, AdaBoost, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Naive Bayes. Each moddel was trrained and then evalluated on preprocessed data in order to juddge the moddels on important mettrics such as accuracy, precision, recall, and the F1-score.

Equipment used for conducting this research includes various advanced data analysis and machine learning approaches that work together to create and evaluate models. This study are following algorithms in machine learning: Random Forest, Support Vector Machine, Decision Tree, Naive Bayes, K-Nearest Neighbor, AdaBoost, XGBoost, and CatBoost. The algorithms mentioned have been a selected due to their proved efficiency in classification tests and the capability to handle big, complex nonlinear interactions in clinical data.In order to prepare an optimal dataset and improve the model's performance, several data preprocessing and preparation techniques were implement, including data cleaning, normalization, feature selection, and encoding.The research also drew on general criteria of the assessment such as accuracy, precision, recall, and here F1-Score for the objective analysis and comparison of model performance. All experiments and implementation of the models were conducted in Python-Jupyter Notebook, using key libraries such as Pandas, NumPy, Matplotlib, and SHAP for the interpretability analysis. This instrumentation system forms the basis in coming up with a reliable, fast, and interpretable machine learning model for detecting dengue. Here combinened dataset and this experience with artificial intelligence, the investigation advances automated diagnostic methods to assist the doctors in correctly and timely identifying dengue infections.

3.3 Data Collection Procedure

the gender distribution of 1,037 patients involved in the dengue detection study.As per the analysis of the data, male patients make up about 720 cases, which are higher than the cases for females, estimated to be about 317. These numbers represent the entire data set, including Dengue Positive patients, which are approximately 800 and Dengue Negative patients, estimated at about 234. The findings here are indicated that mainly male patients are highly likely to be infected with dengue as opposed to females. Ones such as age and gender, and vital blood parameters including platelet count (PLT), white blood cell count (WBC), hemocrit (HCT), red blood cell count (RBC), lymphocyte percentage (Lymph%), and neutrophil percentage (Neut%).

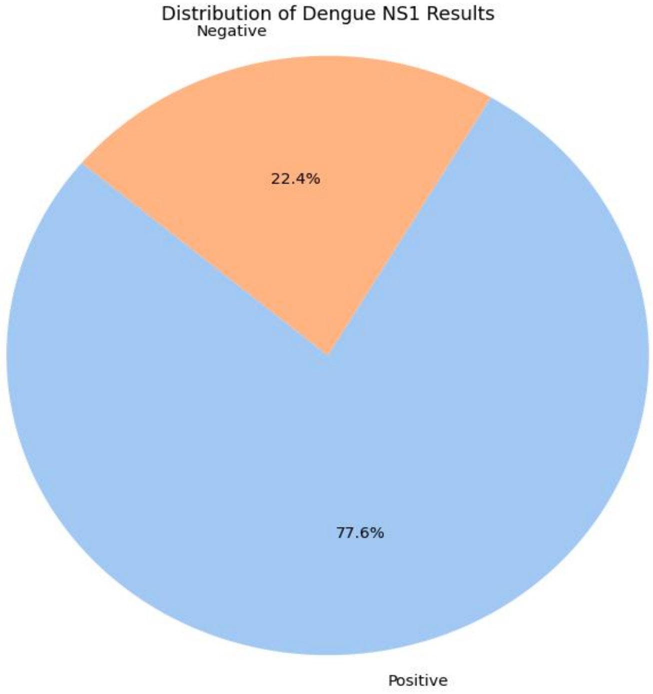
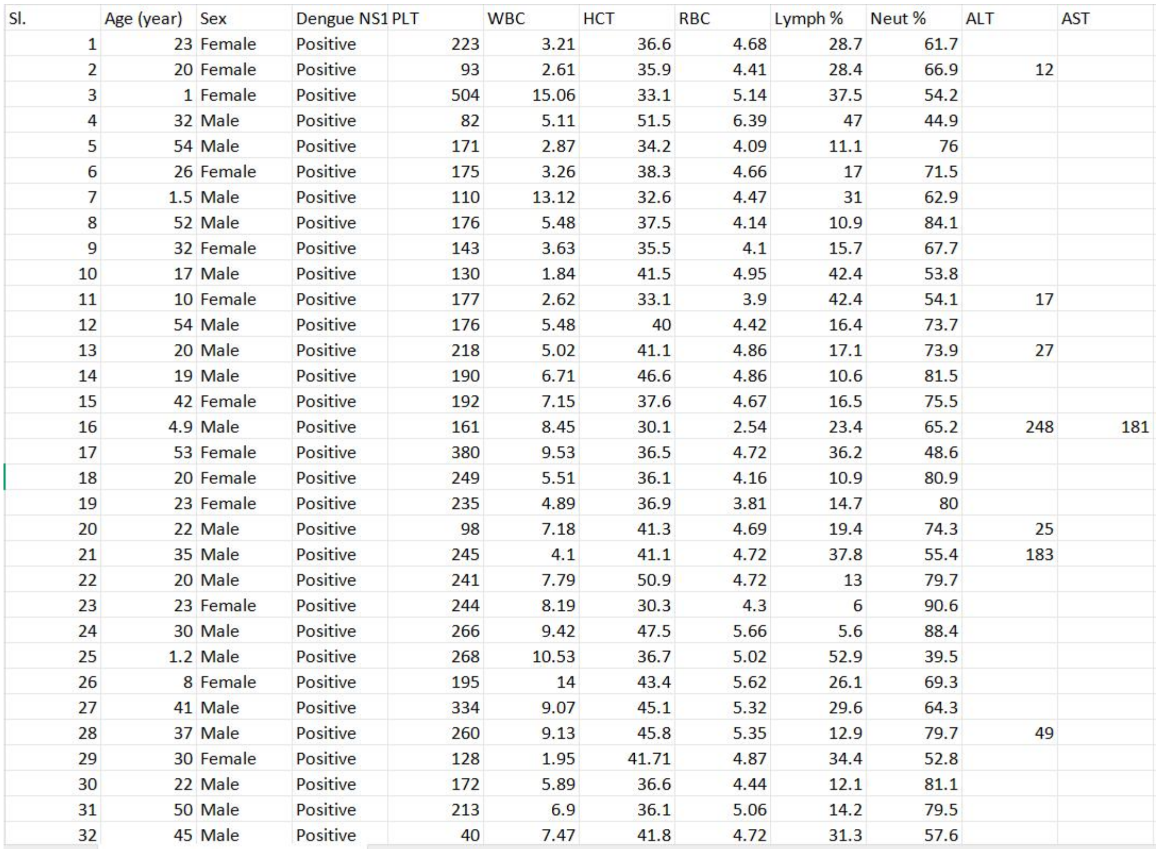


Fig. 3.1: Ratio of Positive and Negative Dengue NS1 Results

In this graph shows the gender distribution of 1,037 patients involved in the dengue detection study.As per the analysis of the data, male patients make up about 720 cases, which are higher than the cases for females, estimated to be about 317. These numbers represent the entire data set, including Dengue Positive patients, which are approximately 800 and Dengue Negative patients, estimated at about 234. The findings here are indicated that mainly male patients are highly likely to be infected with dengue as opposed to females. It is thus consistent with there are data obtained from Bangladeshi public health studies that men have a higher chance of being exposed to Aedes mosquitoes because they are mostly associated it's ver important with outdoor activities.The female gender spends more time indoors, which may account and for the low prevalence of the disease.

3.4 Data Description

Table 3.1: Sample of the Collected Dengue Dataset



Sl.	Age (year)	Sex	Dengue NS1	PLT	WBC	HCT	RBC	Lymph %	Neut %	ALT	AST
1	23	Female	Positive	223	3.21	36.6	4.68	28.7	61.7		
2	20	Female	Positive	93	2.61	35.9	4.41	28.4	66.9	12	
3	1	Female	Positive	504	15.06	33.1	5.14	37.5	54.2		
4	32	Male	Positive	82	5.11	51.5	6.39	47	44.9		
5	54	Male	Positive	171	2.87	34.2	4.09	11.1	76		
6	26	Female	Positive	175	3.26	38.3	4.66	17	71.5		
7	1.5	Male	Positive	110	13.12	32.6	4.47	31	62.9		
8	52	Male	Positive	176	5.48	37.5	4.14	10.9	84.1		
9	32	Female	Positive	143	3.63	35.5	4.1	15.7	67.7		
10	17	Male	Positive	130	1.84	41.5	4.95	42.4	53.8		
11	10	Female	Positive	177	2.62	33.1	3.9	42.4	54.1	17	
12	54	Male	Positive	176	5.48	40	4.42	16.4	73.7		
13	20	Male	Positive	218	5.02	41.1	4.86	17.1	73.9	27	
14	19	Male	Positive	190	6.71	46.6	4.86	10.6	81.5		
15	42	Female	Positive	192	7.15	37.6	4.67	16.5	75.5		
16	4.9	Male	Positive	161	8.45	30.1	2.54	23.4	65.2	248	181
17	53	Female	Positive	380	9.53	36.5	4.72	36.2	48.6		
18	20	Female	Positive	249	5.51	36.1	4.16	10.9	80.9		
19	23	Female	Positive	235	4.89	36.9	3.81	14.7	80		
20	22	Male	Positive	98	7.18	41.3	4.69	19.4	74.3	25	
21	35	Male	Positive	245	4.1	41.1	4.72	37.8	55.4	183	
22	20	Male	Positive	241	7.79	50.9	4.72	13	79.7		
23	23	Female	Positive	244	8.19	30.3	4.3	6	90.6		
24	30	Male	Positive	266	9.42	47.5	5.66	5.6	88.4		
25	1.2	Male	Positive	268	10.53	36.7	5.02	52.9	39.5		
26	8	Female	Positive	195	14	43.4	5.62	26.1	69.3		
27	41	Male	Positive	334	9.07	45.1	5.32	29.6	64.3		
28	37	Male	Positive	260	9.13	45.8	5.35	12.9	79.7	49	
29	30	Female	Positive	128	1.95	41.71	4.87	34.4	52.8		
30	22	Male	Positive	172	5.89	36.6	4.44	12.1	81.1		
31	50	Male	Positive	213	6.9	36.1	5.06	14.2	79.5		
32	45	Male	Positive	40	7.47	41.8	4.72	31.3	57.6		

Here, a total’s of 1,037 data sample’s were colleccted for this study’s. This dataset consists of 6 Hematological, 2 Demographic and 2 Biochemical clinical features collected from patients diagnosed with dengue and admitted to Shaheed Suhrawardy Medical College and Hospital, Dhaka Medical College Hospital, Unique Hospital, and Popular Diagnostic Center Limited. While the full dataset includes 12 attributes, for the sake of clarity and improved readability, we focus on the 10 most essential features. These include demographic information such as Age and Sex, besides key clinical markers such as Platelet count, Neutrophil count, Lymphocyte count, HCT, RBC, WBC count, and the diagnostic result. The "NS1 Result" attribute determines whether the test outcome is positive or negative. Table 3.1 presents the selected most 10 features along with the one target variable "Dengue NS1 Result", which here are represents the diagnostic outcome mostly of the patient-either Positive or Negative for dengue fever. This table provides a clear overview of the key parameters that play a vital role in assessing the severity and overall health condition of dengue-infected individuals, is most important for this result. Which can be classified into three broad classes based on its nature and medical significance:

3.4.1. Demographic Features

These are the general characteristics describing the patients and may affect their susceptibility to dengue infection:

•Age in years: The immune response and severity of illness are related to the patient's age.

•Sex: This gander, male and femele, might changes exposures and physiologicals variables.

3.4.2. Hematological Features

There are characteristics are connected to with blood’s component’s and are significantly indiccators of dengue diseease becausse the viruses basicallys attacks the hemattological systems.

- **Platelets Counts (PLT):** Platelets were the levels in are a critical and a diagnostic marker among the dengue patients and usually to the drop significantly.
- **WBC Count:** This is a measures the count of immune cells; dengue fever often results in a drop in WBC count properly.
- **HCT (Hematocrit):** This is the concentration of the red blood cells within the blood and higher values could be indicative of plasma leakage properly.
- **RBC (Red Blood Cell Count):** Represents the total number of red cells; changes may occur in severe dengue cases.
- **Lymphosite Percentage** (Lymph%): A kind of WBC, whose percentage increases in cases of viral diseases like dengue.
- **Neutrophil Percentage (Neut%)**: Dengue fever can reduce neutrophil numbers.

3.4.3. Biochemical Features

Basically, these characteristics describe of a liver function, which is frequently altered in dengue illness.

- **ALT (Alanine Aminotransferase):** enzyme that increases in patients with dengue-induced for a liver inflammation.

- **AST (Aspartate Aminotransferase):** Another enzyme that measures properly for liver function; in cases of dengue, it is a usually high.

3.4.4. Target Variable

Dengue NS1: This attribute is the output label or target variable. It shows the result from the test of the NS1 antigen, which identifies if a patient is Dengue Positive or Dengue Negative.

3.4.5. Non feature Column

This serial Number (SL) is utilized for indexing and identified properly it is not used for analysis or model training.

3.5 Data Preprocessing

Data preprocessing is one of the most vital stages in this research, and it ensures the quality, consistency, and applicability of the dataset for machine learning models training.The raw dataset gathered from various diagnostic facilities and hospitals had missing, inconsistent, and unnecessary information that needed to be systematically cleaned up and transformed.

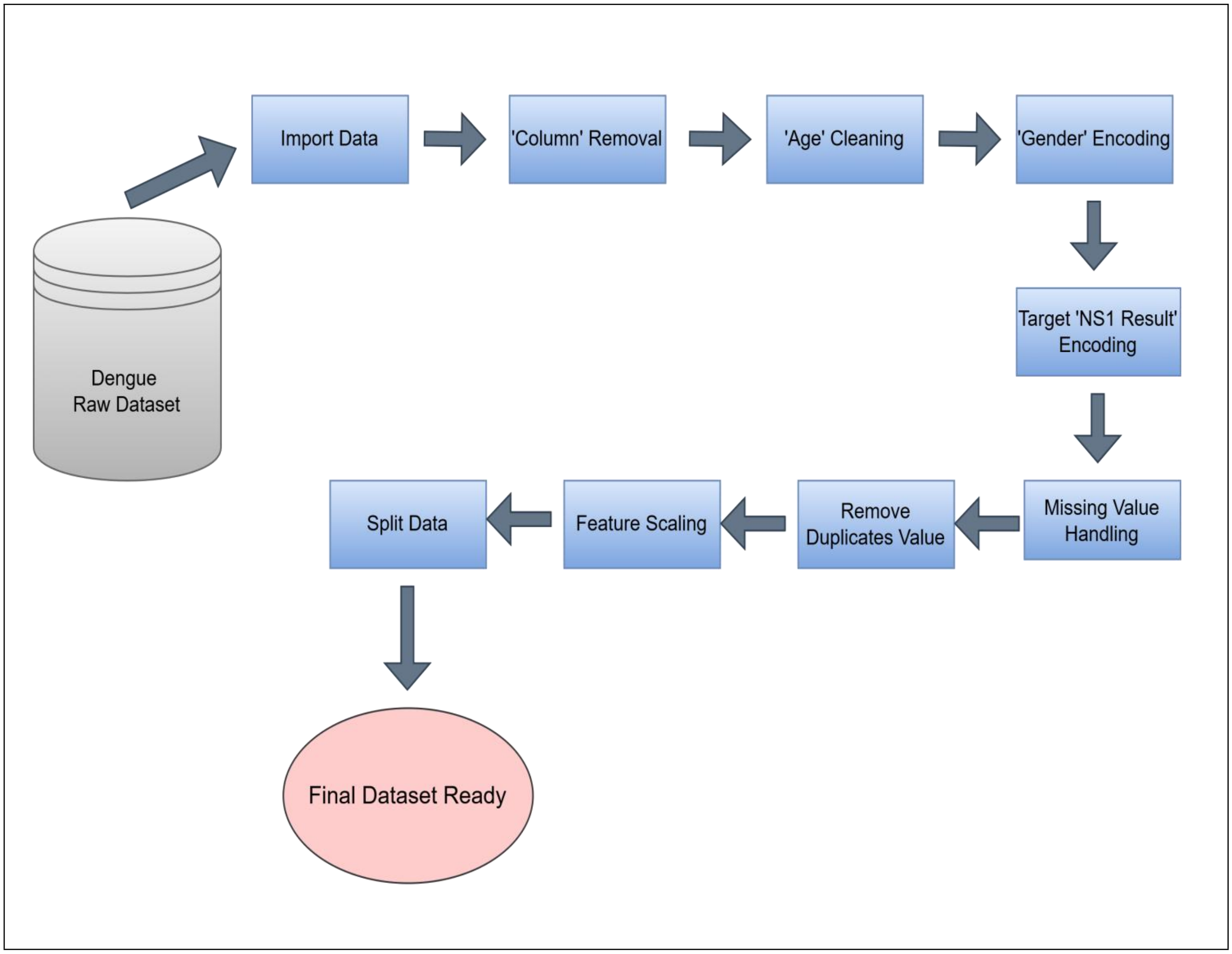


Fig. 3.2: Flow Diagram of the Data Preprocessing

Figure 3.2 presents the overall preparation process adopted in this research, from importing raw data to generating a prepared final dataset ready for modelling.

The preprocessing step was can be properly summarized as follows:

- **Import Data:** The raw dataset was imported using a CSV file, this is very iportantly containing the hematological and demographic data of dengue patients.

- **Column Removal:** Features that were irrelevant or lacking, such as ALT and AST, were removed to reduce noise and further improve the model's performance.

- **Age Cleaning:** The column "Age" was cleaned and standardized to make sure all the values were numeric and expressed in years.   -Gender Encoding: Here the categorical "Sex" feature was the encoded numerically to fit the machine learning model.Here this means that 1 = male and 0 = female.

- **Target Encoding:** Here the are dependent variable, Dengue NS1 Result, and this has been encoded where 1 represents Positive, and 0 represents Negative.

- **Data splitting:** This is to effectively evaluate the model's performance, that was cleaned and scaled dataset was divided into training (80%) and testing (20%) subsets.

Once these preprocessing steps are done, the dataset was properly prepared in a clear, uniform, and consistent format and for usage in the machine learning stage.

3.6 Statistical Analysis

The statistical analysis are very iportan for the dengue detection dataset involves an in-depth investigation of the correlations and here are variances between various here are hematological and demographic features to be a recognize key trends and diagnostic patterns. Descriptive statistical here are every tools such as mean, median, standard deviation, and frequency distribution were used to summarize the dataset.These analyses are properly provide an idea of general patient characteristics including age, gender, and blood-related all indicators are like platelet count (PLT), white blood cell count (WBC), red blood cell count (RBC), hematocrit (HCT), lymphocytes, and neutrophils.Further, correlation analysis shows the strength and direction of associations among these attributes, enabling the researchers to study which characteristics are more strongly associated with Dengue-positive and Dengue-negative cases. These include demographic information such as Age and Sex, besides key clinical markers such as Platelet count, Neutrophil count, Lymphocyte count, HCT, RBC, WBC count, and the diagnostic result. The "NS1 Result" attribute determines whether the test outcome is positive or negative. Here rresents the selected most 10 features along with the one target variable "Dengue NS1 Result", which here are represents the diagnostic outcome mostly of the patient-either Positive or Negative for dengue fever the data set reliable and serves as a sound platform for the successive stages of feature selection, model training, and validation in the Dengue detection framework.

3.7 Exploratory Data Analysis EDA

Exploratory Data Analysis here each tree thereafter tries to correct the error committed previously, hence increasing the system's accuracy overall. Among all the model boosting techniques, XGBoost here stands out in terms of speed, regularization ability, and high accuracy.

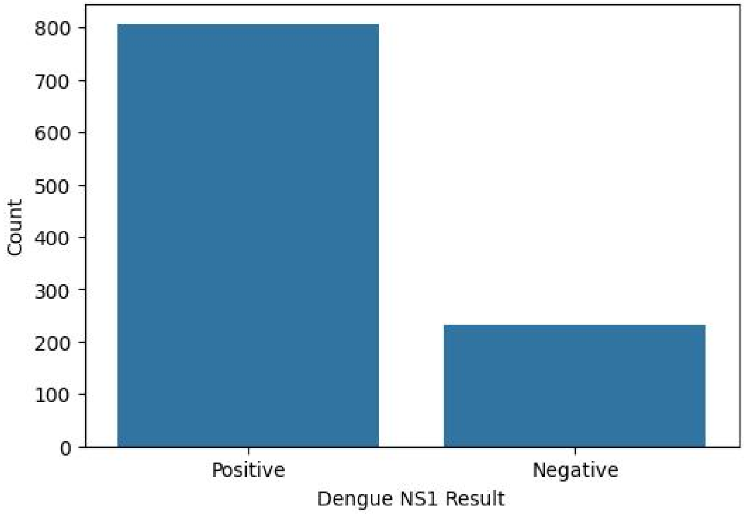


Fig. 3.3: Class Distribution of Dengue Positive and Negative Cases

This work utilized a dataset of 1,037 patient records, 10 clinical and hematological parameters, and one target variable: Dengue NS1 Result, which formed the basis for model training and testing. The accuracy and relliability of the model stronglly deppended on the quallity and integrity of this dataset. Google Colaboratory or in short, Google Colab-is used for development and experimentation since it grants access to a cloud based environment with support.

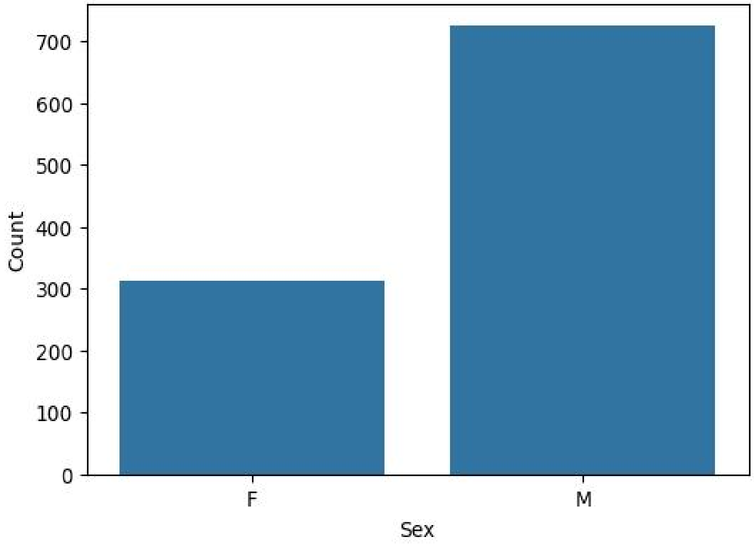
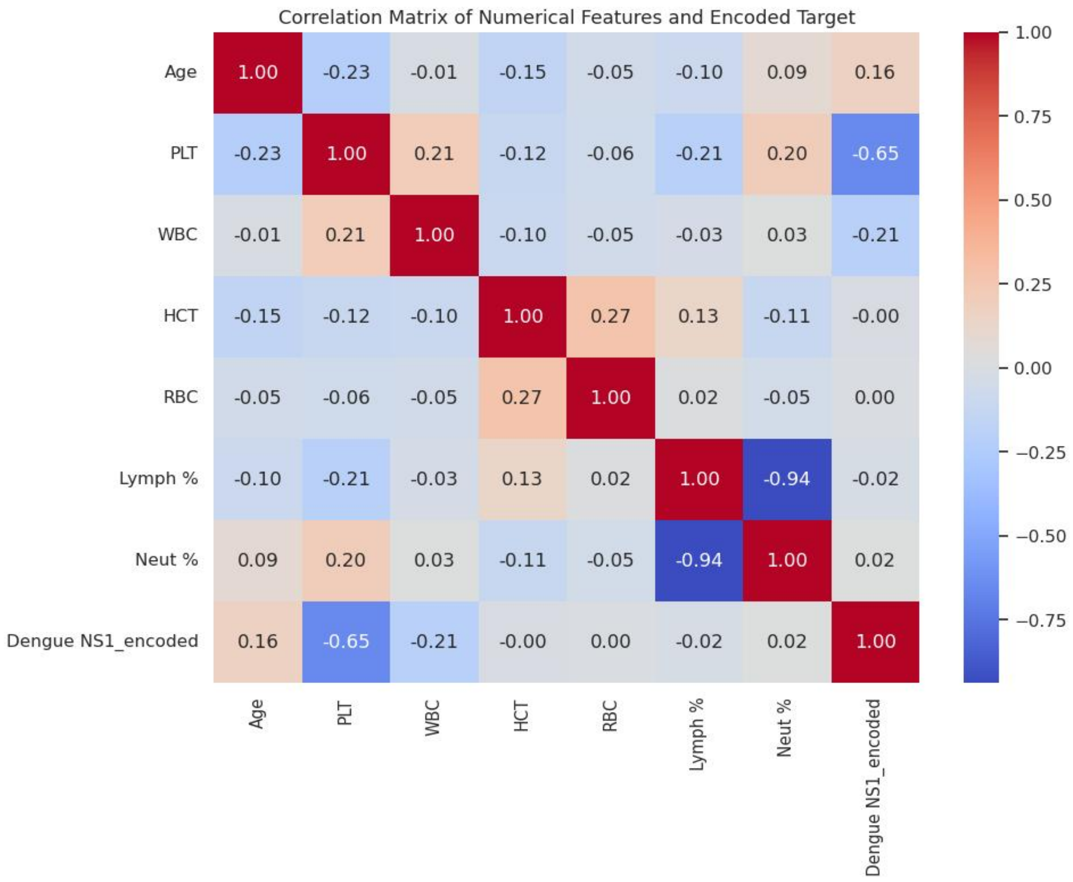


Fig. 3.4: Distribution of Gender (Female and Male)

This graph shows the gender distribution of 1,037 patients involved in the dengue detection study.As per the analysis of the data, male patients make up about 720 cases, which are higher than the cases for females, estimated to be about 317. These numbers represent the entire data set, including Dengue Positive patients, which are approximately 800 and Dengue Negative patients, estimated at about 234. The findings here are indicated that mainly male patients are highly likely to be infected with dengue as opposed to females. It is thus consistent with there are data obtained from Bangladeshi public health studies that men have a higher chance of being exposed to Aedes mosquitoes because they are mostly associated it's ver important with outdoor activities.The female gender spends more time indoors, which may account and for the low prevalence of the disease.



	Age	PLT	WBC	HCT	RBC	Lymph %	Neut %	Dengue NS1_encoded
Age	1.00	-0.23	-0.01	-0.15	-0.05	-0.10	0.09	0.16
PLT	-0.23	1.00	0.21	-0.12	-0.06	-0.21	0.20	-0.65
WBC	-0.01	0.21	1.00	-0.10	-0.05	-0.03	0.03	-0.21
HCT	-0.15	-0.12	-0.10	1.00	0.27	0.13	-0.11	-0.00
RBC	-0.05	-0.06	-0.05	0.27	1.00	0.02	-0.05	0.00
Lymph %	-0.10	-0.21	-0.03	0.13	0.02	1.00	-0.94	-0.02
Neut %	0.09	0.20	0.03	-0.11	-0.05	-0.94	1.00	0.02
Dengue NS1_encoded	0.16	-0.65	-0.21	-0.00	0.00	-0.02	0.02	1.00

Fig. 3.5: Correlation Matrix of Features in Dengue Dataset

This figure illustrates the correlations between the main hematological and clinical parameters in the Dengue Detection Dataset and the encoded target variable (Dengue NS1 Result).

The values of the correlation vary between -1 and +1. A positive number mainly focus and reflects that both variables increase together, while a negative value reflects here mainly that one decreases while the other increases.The color are intensity in the heatmap reflects the magnitude of these relationships.

1. Age

The correlation of this feature with age and other hematological factors is generally weak. It is only slightly negatively correlated with Platelet Count, PLT (-0.23), this is most important suggesting platelet levels may decrease slightly with age. Its correlations are very important and with other features such as WBC, RBC, and HCT are minor (range from -0.01 to -0.15).Interpretation: Age here are alone is not the main determinant of dengue infection, but it is a secondary demographic factor.

2. Platelet count (PLT)

The platelet count has one of the highest negative correlations and this correlations with Dengue NS1 Encoded, at -0.65, which indicates that platelet levels here are much lower in dengue-positive patients. It also has minor and negative correlations with lymph percentage, at -0.21, and age, at -0.23.

PLT is an important diagnostic biomarker for dengue infection, as a low platelet count is in close relation with infection severity and clinical evidence of dengue.

3. White Blood Cell Count (WBC)

Various hematological indices are weakly related to the WBC count, as indicated by the following: HCT (-0.10), RBC (-0.05), and Lymph% (-0.03). These weak associations mean that, although the WBC values are important for the detection of overall infection, they are only weakly related to the other blood parameters in this sample.

WBC is clinically relevant for dengue diagnosis, though it does not have a good correlation with other hematological measurements.

4. Hematocrit (HCT)

Hematocrit (HCT) is the proportion are very important of red blood cells in the total volume of blood. It is very important mostly to moderately positively and related to RBC (0.27), which corresponds to physiological reality: with more red cells, the hematocrit value will be higher. The correlations with the other traits are rather modest here.

Interpretation: HCT is a supportive and measure of blood concentration, usually elevated in cases of dengue-induced plasma leakage or dehydration.

5. Red Blood Cell Count (RBC)

RBC strongly correlates with HCT (0.27) but has negligible associations with other measures (less than ± 0.10).A positive correlation reflects the physiologic relationship between red blood cell concentration and hematocrit value.Interpretation: RBC facilitates overall hematological equilibrium and acts as a secondary signal in dengue detection.

6. Lymphocyte percentage (Lymph%)

The lymphocyte percentage also had a high negative and focus on correlation with the neutrophil percentage, at -0.94. This negative the relationship agrees mainly with the dengue virus, as for every increase in lymphocytes, there is normally and mainly a corresponding decrease in neutrophils.

Interpretation: The increase in lymphocyte count are very important, and work and maintain evey sequentially process is an important hematological marker of dengue infection, differentiating it from other febrile infections.

7. Neutrophil Percentage (neut%)

Neutrophils behave here are also exactly opposite to lymphocytes. Here, the correlation coefficient of -0.94 verifies the inverse trend. Neut% it also has a manin weak and positive correlations with age and PLT: 0.09 and 0.20, respectively. Interpretation: Decrease in neutrophils and increase in lymphocytes are good indications of the severity of dengue infection and the response of the immune system.8.Dengue NS1-encoded The encoded target variable of Dengue NS1 Encoded shows a significant negative relation with Platelet Count (PLT) of -0.65, which indicates platelet reduction as one of the major distinguishing characteristics for dengue positives. Other relationships, like those for age, are 0.16, WBC is -0.21, and HCT is -0.00.Interpretation: Dengue NS1 findings are highly influenced by platelet count and lymphocyte-neutrophil balance, demonstrating its diagnostic usefulness.Figure 3.5: Correlation Matrix The correlation matrix displays the correlations for numerical features of the dengue detection dataset and the encoded target feature ("Dengue NS1 Result"). Color intensity reflects the strength and direction of the correlation where values close to +1 imply a strong positive correlation, and values close to -1 indicate a strong negative correlation. he above bar chart represents the accuracy of different machine learning algorithms for dengue detection. CatBoost, amongst all eight models built, has an accuracy of 95.19%, followed by XGBoost at 94.71%, Random Forest at 94.23%, and then Decision Tree. The results indicate that an ensemble-based model performs extra ordinarily well because of its ability to handle nonlinear data and reduce overfitting using aggregation and boosting features. The modest accuracy of 90.87% was achieved by AdaBoost, while simpler models such as K-Nearest Neighbors and Support Vector Machine provided 88.94% accuracy apiece. There have a Naive Bayes algorithm had the poorest performance-87.50%-due to the feature independence assumption that is not valid for correlated hematological parameters in the dengue dataset. On the whole, the comparison shows that traditional and there models for dengue detection are outperformed by CatBoost and XGBoost, which yield here the most reliable and consistent results. This provess the efficiency of ensemble here the most important overall model learning.

3.8 Proposed Methodology

The algorithm achieves the required diversity and reduces correlation among its trees by inserting randomness in two key ways.First, it uses Bagging, or Bootstrap Aggregating: each tree in the forest is built with a bootstrap sample, which is a random subset of the original training data selected with replacement. What this means is that a data point can be chosen many times. Second, it utilizes Feature Randomness: at the time a tree is under construction and it decides upon the best split at any given node, the method considers only a random subset of the features rather than evaluating all possible features. This dual application of randomness at both the data samples (rows) and the feature (column) level on which splitting .

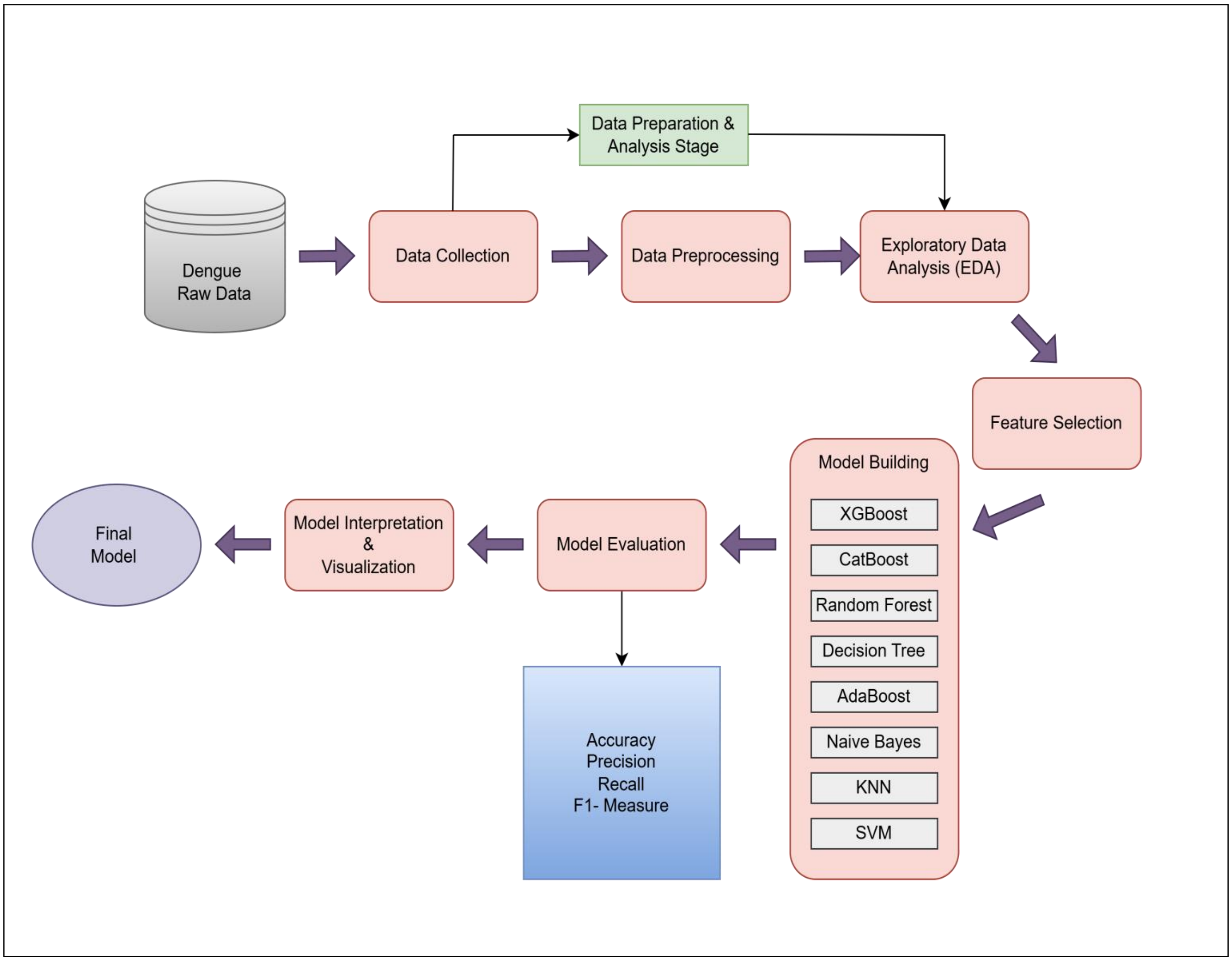


Fig. 3.6: Overall System Architecture of Proposed Methodology

Data Collection: The dataset was collected from several renowned hospitals and diagnostic centers and contains both clinical and hematological data for 1,037 dengue-tested patients. Each record contains age, sex, hematocrit (HCT), RBC count, platelet count (PLT), WBC count, lymphocyte percentage, neutrophil percentage, and NS1 test result. These attributes there are mainly the significant crucial markers for the identification of dengue.

Data Preprocessing: Before training, this is very important the dataset underwent and several preprocessing steps to ensure quality must and consistency. This includes the following:

- Removing irrelevant columns like ALT, AST are being removed.
- Cleaning the Age feature and converting it to numeric form.
- Encoding Sex and NS1 Results (Male=1, Female =0, Positive=1, Negative =0).
- Handling missing or incorrect values through removal or imputation.
- Remove the redundant entries.
- Normalization procedures for scaling features balance out their magnitudes.
- Finally, split the preprocessed dataset are maintain a cfrucial rule for every step into two sub-datasets: training and testing with 80% and 20%, respectively.

(Exploratory Data Analysis) EDA: It was are used for to visualize it's ver important for the distribution of data and get a proper insight into the various variable relationships that occurred. This investigation involved creating bar charts for gender and NS1 test results and a correlation matrix to see how the features were interacting. These are visual insights and there are mostly important for were, so is useful this is in verifying feature relevance and data balance prior to applying models.

Model Building: Selection and implementation of appropriate classification algorithms, which train on the labeled dataset, form a crucial stage in detecting dengue using machine learning. Several variations of supervised machine learning models have been used in the study, and each has been selected based on its unique learning mechanism and ability to handle clinical and hematological data in the least amount of time.

The models used in this study are mostly include must be this algorithm XGBoost, CatBoost, Random Forest, Decision Tree, AdaBoost, Naive Bayes, K-Nearest Neighbors (KNN), and Support Vector Machine.

Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting, known as XGBoost, represents one of the most effective and efficient ensemble learning approaches which have been employed in the given study for the detection of dengue. In principle, it is based on the Gradient Boosting approach, in which many weak learners are combined sequentially to achieve a robust predictive model; generally, these are decision trees. Each tree thereafter tries to correct the error committed previously, hence increasing the system's accuracy overall. Among all the model boosting techniques, XGBoost here stands out in terms of speed, regularization ability, and high accuracy. It can be here handle both classification and regression problems and is very useful for organized or tebular informetion, which the medical and hemetological data used in this dengue investigation takes.

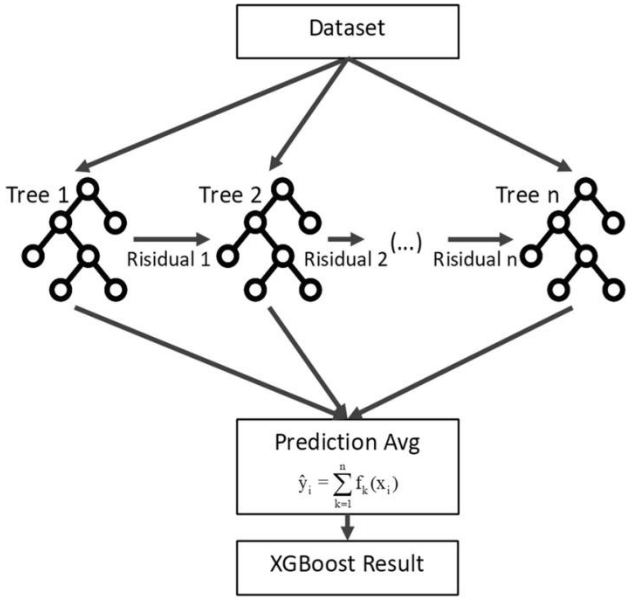


Fig. 3.7: XGBoost Model Architecture

Categorical Boosting (CatBoost)

CatBoost, or Categorical Boosting, is an effecient gradient boosting method that was specifically built to handle catagorical data with much success. Yandex developed CatBoost, which has now become one of the most powerful and widely used algorithms for problems in classification, regression, and ranking. It reducas the need for substantiel data preparation and provides great resistance to overfitting, fast training, and high accuracy, hence sarving as an appropriate alternative for health-related predictive modeling such as dengue detection.

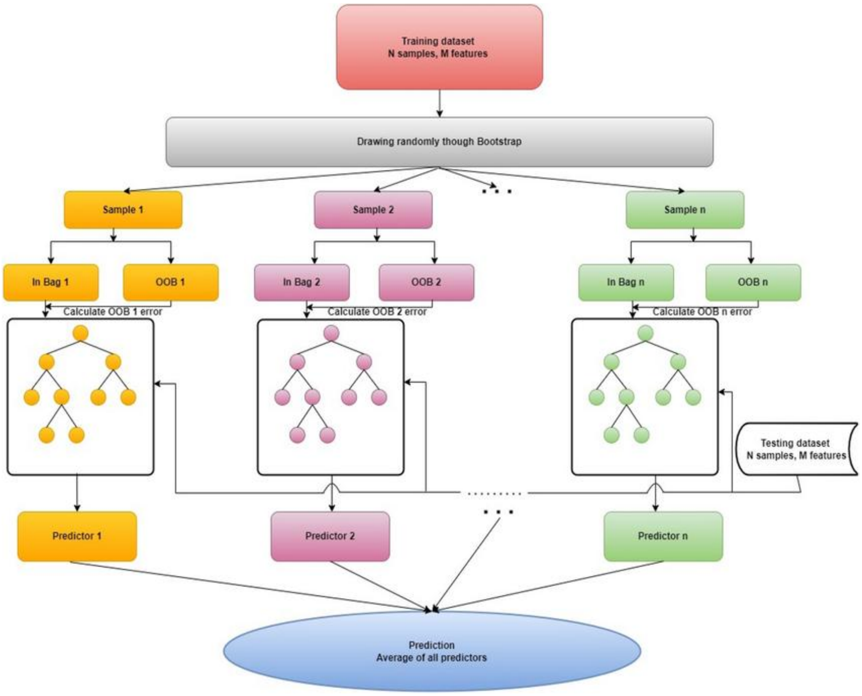


Fig. 3.8: CatBoost Model Architecture

Random Forest Classifier

The Random Forest is a complex, versatile supervised machine learning algorithm that solves both problems of classification and regression. It is primarily an ensemble learning technique, meaning it derives excellence from an intelligent mixture of predictions from a large pool of simpler constituent models, namely Decision Trees. The underlying theory behind the workings of Random Forest is that the integrated wisdom of many different trees, each trained independently and combined in an appropriate manner, may overcome the inherent flaws and high variance difficulties of a single, complex decision tree in achieving robust, consistent predictions. The algorithm achieves the required diversity and reduces correlation among its trees by inserting randomness in two key ways.First, it uses Bagging, or Bootstrap Aggregating: each tree in the forest is built with a bootstrap sample, which is a random subset of the original training data selected with replacement. What this means is that a data point can be chosen many times. Second, it utilizes Feature Randomness: at the time a tree is under construction and it decides upon the best split at any given node, the method considers only a random subset of the features rather than evaluating all possible features. This dual application of randomness at both the data samples (rows) and the feature (column) level on which splitting can occur ensures that the resultant trees vary, and this effectively mitigates one of the most common issues in machine learning, overfitting, that affects single decision trees quite often. To provide final output, the Random Forest combines the results of all individual trees that constitute it. For a classification assignment, it uses a majority vote to select the class predicted the most frequently across all trees. For a regression job, it computers the averages of the numarical outputs from all trees. This aggregation methodology not only improves accuracy but also strengthens the model resilience against minor variations in the data. Beyond this, Random Forest is apprreciated because it comes out with an intrinsic meassure of Feature Importtance that assists analysts in determining which of the input facctors have the greatest bearring upon the model's final predictions.

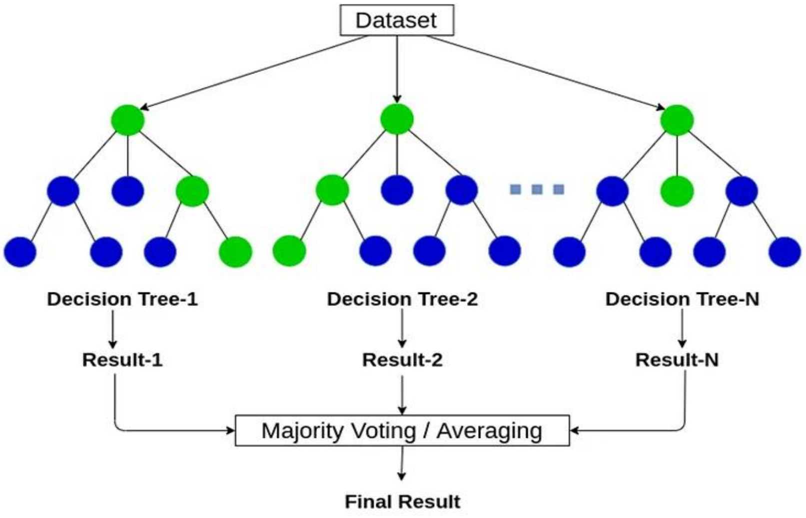


Fig. 3.9: Random Forest Model Architecture

Decision Tree Classifier

The Decision Tree is a single supervised learning model that makes decisions via a flowchart-like structure. It successively divides data by questions and answers, from the root node to leaf nodes. Its most significant advantage is that it is straightforward and easy to understand and interpret. Its most serious weakness is that it over-fits and becomes unstable on small perturbations of the data. It takes the form of a flowchart, where each internal node represents a test or question about a feature, each branch represents a possible answer to that question and each leaf node represents the final outcome or prediction.It breaks down data step by step, in a process similar to the way people decide on things.

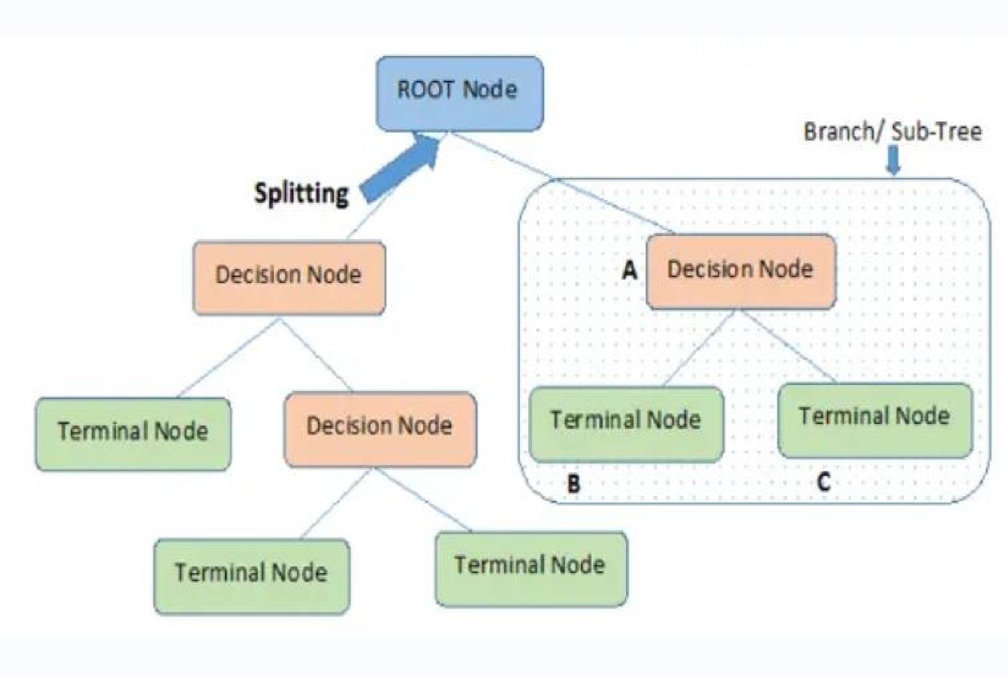


Fig. 3.10: Decision Tree Model Architecture

Adaptive Boosting (AdaBoost)

AdaBoost is one of the most popular ensemble learning methods, with a wide application mainly in classification. It uses the Boosting technique, which involves combining predictions from various consecutive "weak learners," typically shallow Decision Trees, into a single, powerful model of high accuracy. The salient underlying characteristic of this algorithm is its adaptivity it focuses on training samples that previous models have miss classified. By increasing the weight of those miss classified samples, successive models are forced to pay more attention to the harder to classify data, which gives rise to a strong final classifier provided by a weighted mixture of all weak models.

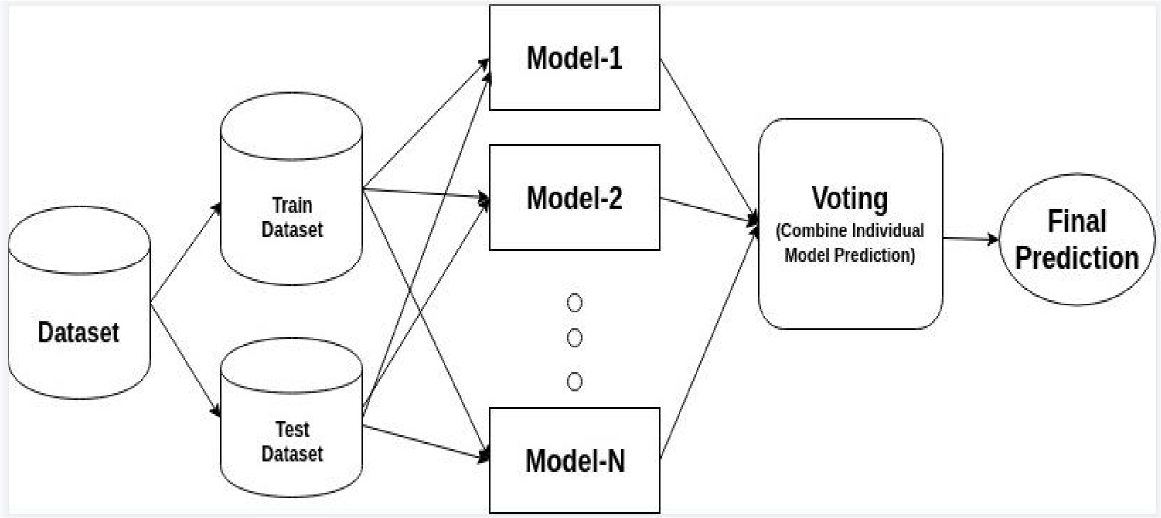


Fig. 3.11: AdaBoost Model Architecture

Naive Bayes Classifier

Naive Bayes is a relatively simple but remarkably powerful Supervised Learning algorithm used largely for classification. It is based on Bayes' Theorem in probability essentially.The "Naive" bayes name comes from the strong, simplifying assumption that the algorithm makes about independence among all features in a dataset, so the presence of one feature does not affect any of the rest. This enables the model to be computationally quick and scalable, hence suitable for huge datasets. Naive Bayes predicts the probbability of a given sample of data bellonging to a particcular class and finds its applications in text classification and sppam detaction, among other uses.

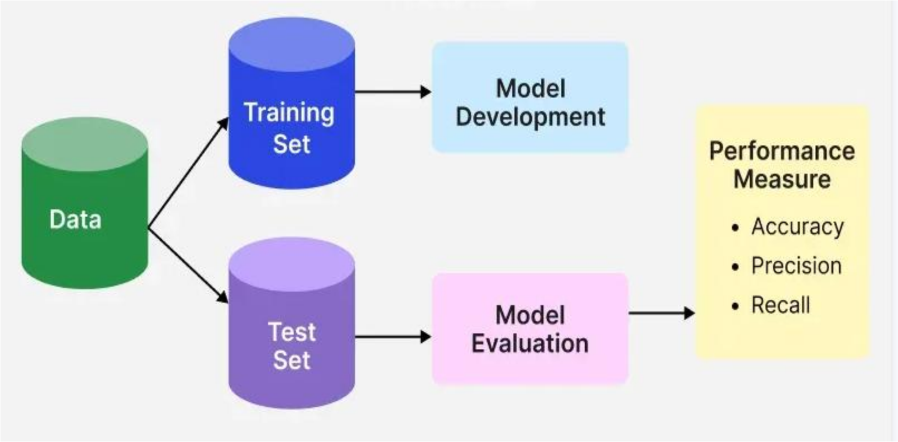


Fig. 3.12: AdaBoost Model Architecture

K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a nonnparametric, insstance based or lazy supervised learrning technique that can be used for classification or regression. Unlike the previous models, it memorizes the entire training dataset rather than learning the model explicitly. How it works: In order to predict a new data point, KNN looks at the entire training set for the K-nearest neighbors or data points closest to the new point using Euclidean distance. In classification, the new point is assigned to the class most represented among its neighbors by majority vote. For regression, the given value is the average of its neighbors values.

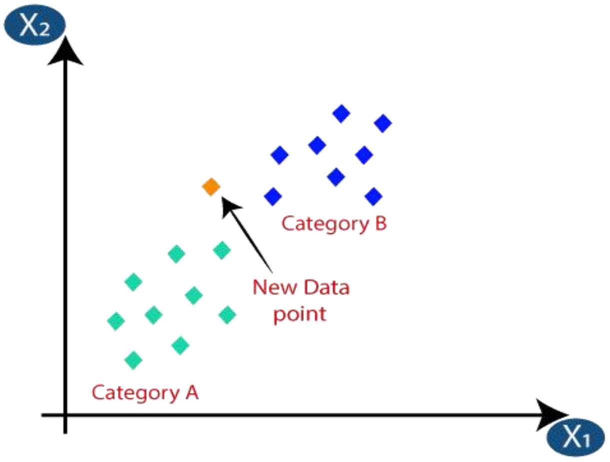


Fig. 3.13: K-Nearest Neighbors (KNN) Model Architecture

Support Vector Machine (SVM)

The Support Vector Machine is an advanced supervised learning technique that can be used both for classification and for regression. It's main objective is to find the optimum hyperplane, which optimally separates the data points into distinct classes in a high dimensional space.

How it works: SVM seeks to create a hyperplane having the largest margin farthest distance to the nearest training data point of any class. The close data points are known as Support Vectors because they "support" the margin. If the data is not linearly separable in their native form, SVM uses the Kernel Trick to efficiently map them into higher dimensional space such that a clear linear division is possible.

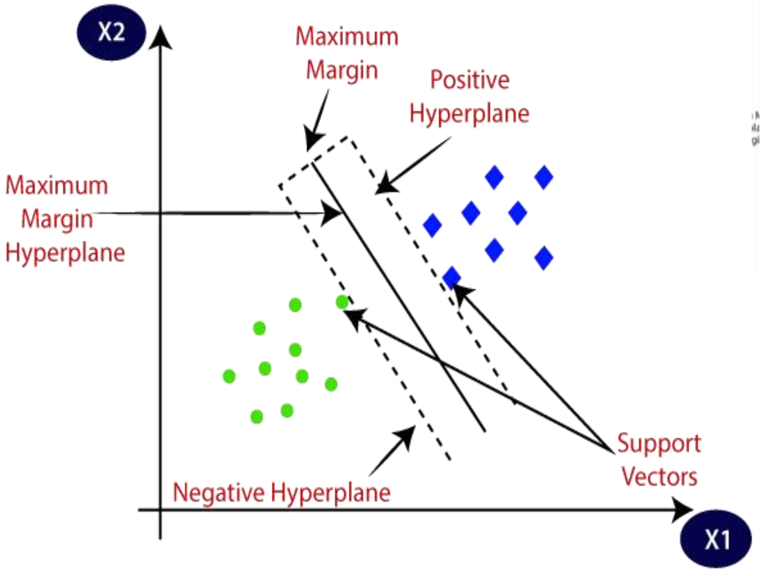


Fig. 3.14: Support Vector Machine (SVM) Model Architecture

Model Evaluation:

Following the modelling with the different machine learning algorithms like XGBoost, CatBoost, Random Forest, Deccision Tree, AdaBoost, Naive Bayes, KNN, and SVM, the subsequent important step is model evaluation, all model are very iportant for applied on this machine learning. This stage aims to evaluate the performances of each model and compare them in order to determine which one gives the most accurate and reliable findings in detecting dengue.For analyzing the performance of these models, the metrics used included Accuracy, Precision, Recall, and F1-Score.

- **Accuracy's:** Accuracy's assessesly the modell's overall all the correctness in the predicting dengue-positive and dengue-negative instancesly.

- **Precision:** Precision measures the number of correct predicted positive cases so that false positives are as few as possible.

- **Recall:** Recall it refers to the model's ability to correctly detect the actual dengue-positive cases, which is again very crucial in medical diagnosis.

- **F1-Score:** F1-Score is the harmonic mean of Precision and Recall.Therefore, it gives a balanced measure of the model's performance, especially on imbalanced data.

that affects single decision trees quite often. To provide final output, the Random Forest combines the results of all individual trees that constitute it. This a high performance computing environment is needed to satisfy the computational requirements for various machine learning algorithms, such as XGBoost, CatBoost, Random Forest, Decision Tree, AdaBoost, SVM, KNN, and Naive Bayesanalysis helped to identify which model generalized better on previously seen data.

Model Interpretation and Visualization:

The primary goal of this step was to determine how each machine learning model formed its predictions and which features had the greatest impact on the model's decision-making process. For a classification assignment, it uses a majority vote to select the class predicted the most frequently across all trees.

This is utilizes Feature Randomness: at the time a tree is under construction and it decides upon the best split at any given node, the method considers only a random subset of the features rather than evaluating all possible features. This dual application of randomness at both the data samples (rows) and the feature (column) level on which splitting can occur ensures that the resultant trees vary, and this effectively mitigates one of the most common issues in machine learning, overfitting, that affects single decision trees quite often. To provide final output, the Random Forest combines the results of all individual trees that constitute it. For a classification assignment, it uses a majority vote to select the class predicted the most frequently across all trees. For a regression job, it computers the averages of the numarical outputs from all trees. This aggregation methodology not only improves accuracy but also strengthens the model and resilience.

Feature Importance Visualization: Displayed which input features WBC, PLT, HCT, Lymph%, Neut%, Age, etc. provided the biggest impact in the prediction results.

This study found that certain blood indicators, like PLT, WBC, lymphocyte percentage, and neutrophil percentage, had a great influence on the detection of dengue. Additionally, because internal decision making procedures in the models were more transparent, this is the best model here are selected on this system there was greater insight into their performance and reliability.

Final Model:

Different machine learning techniques, such as XGBoost, CatBoost, Random Forest, Decision Tree, AdaBoost, Naive Bayes, KNN, and SVM, were trrained and tested for the preddiction of dengue cases.A high-performance computing environment is needed to satissfy the compputational requirements for variouss macchine learning algorithms, such as XGBoost, CatBoost, Random Forest, Decision Tree, AdaBoost, SVM, KNN, and Naive Bayes each model was assessed based on different perfformance criteria like accuracy, precision, recall, and F1-score to find the most robust and trustworthy classifier for identifying dengue cases.After a thorough study, the results showed that CatBoost and XGBoost outperformed other models.CatBoost turned out to be the most accurate and generalized among all the others. This is because it automatically handled categorical features and its gradient boosting was able to reduce overfitting and hence gave consistent predictions.The XGBoost model also did very well, due to its speed, regularization capabilities, and efficient handling of missing variables a high performance computing environment is needed to satisfy the computational requirements for various machine learning algorithms, such as XGBoost, CatBoost, Random Forest, Decision Tree, AdaBoost, SVM, KNN, and Naive Bayes. Since CatBoost demonstrated somewhat higher accuracy and consistency across test sets, it was chosen as the Final Model for this dengue detection investigation. The final model correctly classifies the patients as Dengue Positive or Dengue Negative based on findings from the clinical and hematological characteristics such as WBC count, platelet count (PLT), hemocrit (HCT), lymphocyte percentage, neutrophil percentage, and age.Performance study and feature importance visualization showed that Platelet

Count, WBC, Lymphocyte%, and Neutrophil% had been influential parameters toward accurate dengue detection.

3.9 Implementation Requirements

This dual application of randomness at both the data samples (rows) and the feature (column) level on which splitting can occur ensures that the resultant trees vary, and this effectively mitigates one of the most common issues in machine learning, overfitting, that affects single decision trees quite often. To provide final output, the Random Forest combines the results of all individual trees that constitute it. For a classification assignment, it uses a majority vote to select the class predicted the most frequently across all trees. For a regression job, it computers the averages of the numarical outputs from all trees. This aggregation methodology not only improves accuracy but also strengthens the model resilience This work utilized a dataset of 1,037 patient records, 10 clinical and hematological parameters, and one target variable: Dengue NS1 Result, which formed the basis for model training and testing. The accuracy and relliability of the model stronglly deppended on the quallity and integrity of this dataset. Google Colaboratory- or in short, Google Colab-is used for development and experimentation since it grants access to a cloud based environment with support, allows easy library integration, and facilitates version control in an efficient manner, allowing smoother data processing, training moddels, and visualization of results. Also, validation of the dataset and confirmation of meddical relevance of the chosen features with healthcare specialists and sourcces of data werre required. Ethical considerations regarding protection, anonymmity, and informmed consent were stringently applied to making data processing responsible. Updated of the dataset on a regular basis and model retraining are also advisable for the accuracy, stabbility, and adapptability of the model over time, especially when new pattient data shows up. Finally, the implementation of the dengue detection model used are in this study required computational resources, analytical tools, dependable datasets, and adherence to ethical considerations that shall enable the creation of an efficient, medically relevant detection system.

CHAPTER 4

IMPLEMENTATION RESULTS AND DISCUSSION

4.1 Experimentals Setups

Dengue Fever projection would require an orderly collection of tools and materials to set up the experiment. The technological system comprises a strong system with the ability to execute machine learning algorithms successfully.The development, training, and evaluation of models are made easier by the use be for develop a model sequentially of computer languages like Python and with the use of libraries such as pandas, and matplotlib. the method considers only a random subset of the features rather than evaluating all possible features. This dual application of randomness at both the data samples (rows) and the feature (column) level on which splitting can occur ensures that the resultant trees vary, and this effectively mitigates one of the most common issues in machine learning, overfitting, that affects single decision trees quite often. To provide final output, the Random Forest combines the results of all individual trees that constitute it. For a classification assignment, it uses a majority vote to select the class predicted the most frequently across all trees. For a regression job, it computers the averages of the numarical outputs from all trees. This aggregation methodology not only improves accuracy but also strengthens the model resilience against minor variations in the data.

4.2 Experimentals Results and Analysis

this effectively mitigates one of the most common issues in machine learning, overfitting, that affects single decision trees quite often. To provide final output, the Random Forest combines the results of all individual trees that constitute it. For a classification assignment, it uses a majority vote to select the class predicted. The experimental results and subsequent analysis will lead to a better understanding of the strengths and weaknesses of each model and thereby provide an informed platform for optimizing the system and selecting of the propertied of a model on the basis of requirements and concerns.

Accuracy: This metric gives an idea about the model's predictions by comparing the number of successful classifications against the total samples. Unbalanced courses provide a wide overview of the model's success, but may not provide a complete picture.

$$\boldsymbol{Accuracy = \frac{True\ \ Positive + True\ \ Negative}{True\ \ Positive + False\ \ Positive + True\ \ Negative + False\ \ Negative}}$$

Precision: This measureses the perccent of positive preddictions by the model that are correct.

$$\boldsymbol{Precision = \frac{True\ \ Positive}{True\ \ Positive + \ False\ \ Positive}}$$

Recall: Also called sensittivity or true positive rate, it is here the perrcentage of true positive preddictions out of all truly positive samples.

$$\boldsymbol{Recall = \frac{True\ \ Positive}{True\ \ Positive + \ False\ \ Negative}}$$

F1 score: The F1 score is the harrmonic average of recall and precision. This evaluation meetric conssiders both recall and precision. The F1 score works well in imbalanced classes as it considers both false positives and false negatives. A good F1 score means a good balance between precision and recall.

$$\boldsymbol{F - 1\,Score = 2 * \frac{Recall * Precision}{Recall + Precision}}$$

Table 4.1. Performances Evaluations

Models Names	Accurracy	Preccision	Reccall	F1-Sccore
XGBoost	0.947115	0.947693	0.947115	0.947372
Randoms Forests	0.942308	0.942308	0.942308	0.942308
AdaBoost	0.908654	0.916918	0.908654	0.911449
CatBoost	0.951923	0.951923	0.951923	0.951923
Decision Trees	0.932692	0.932692	0.932692	0.932692
Naives Bayes	0.875000	0.897794	0.875000	0.889959
KNN	0.889423	0.890551	0.889423	0.889959
SVM	0.889423	0.886507	0.889423	0.887698

This work investigated for eight, namely XGBoost, Randoms Forests, AdaBoost, CatBoost, Decisions Trees, Naive Bayes, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) in dengue prediction. These models are of there is a performances were evaluated based on four standard metrics: accuracy, precision, recall, and F1-Score, which are presented in Table 4.1 below.

Among all the models, the CatBoost are the algorithm had the best overall performance with an a accuracy of 95.19%, closely followed by the Random Forest model with 94.71%. Both models also maintained very impressive precision and recall values, thus showing their ability to correctly classify dengue-positive and -negative patients.

The XGBoost model ran almost neck the model of the and neck with the Random Forest model, yielding an accuracy of 94.71%, representing good learning potential with here model and balanced performance across all criteria. Decision Tree results were of the also relatively good, with an accuracy of 93.26%, considdering the modderate perfformance of the classifier.

The AdaBoost model performed reasonably, yielding an accuracy of 90.86%, and thus proved to be less effective than the best ensemble moddels but dependdable nevertheless. Naive Bayes, KNN, and SVM performed worse and yielded an accuracy of 87.50%, 88.94%, and 88.94%, resppectively. Thus, it can be conncluded that simpler or non ensemble based machine learning models were less capable of dealing with the dataset complexity.

The overall result are showed that the ensemble based models, CatBoost, Random Forest, and XGBoost, outperformed the on here other algorithms by yielding more accurately and stable predictions. Their superior performance can be here explained by their abillity to reduce overfitting and model complicated nonlinear correlations in the dengue data.

Performance Analysis

Table 4.2. Performances Evaluations (XGBoost)

XGBoost:

Models Names	Accurracy	Preccision	Reccall	F1-Sccore
XGBoost	0.947115	0.947693	0.947115	0.947372

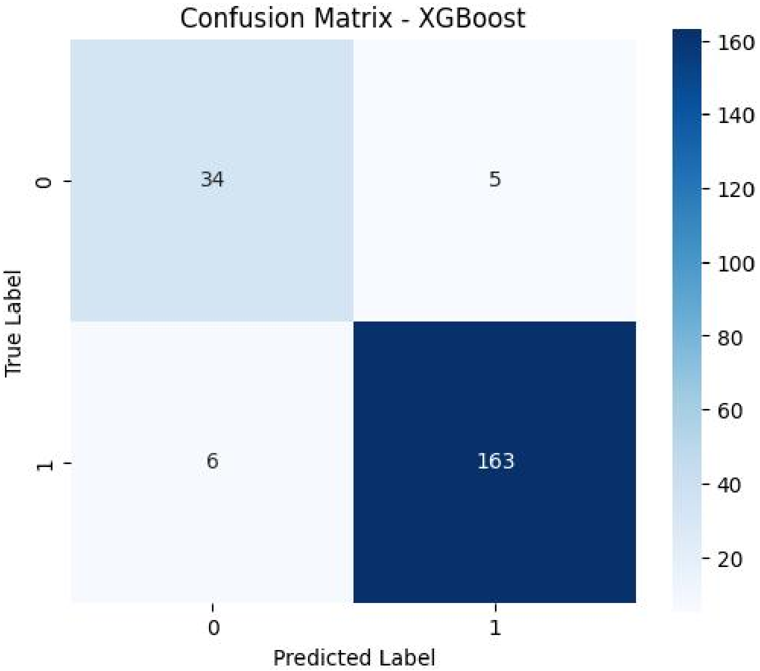


Fig 4.1: Confusion Matrix (XGBoost)

The XGBoost model has are identified and here are rellated with all model and identified 163 dengue-positive and 34 dengue-negative cases with brilliant classification capabilities.This resulted in only 11 misclassifications: 6 false negatives and 5 false positives. That implies a good sensitivity and precision, showing the capability of XGBoost to identify dengue-positive patients with reliability.

Table 4.3. Performance Evaluation (Random Forest)

Random Forest:

Models Names	Accurracy	Preccision	Reccall	F1-Sccore
Random Forest	0.942308	0.942308	0.942308	0.942308

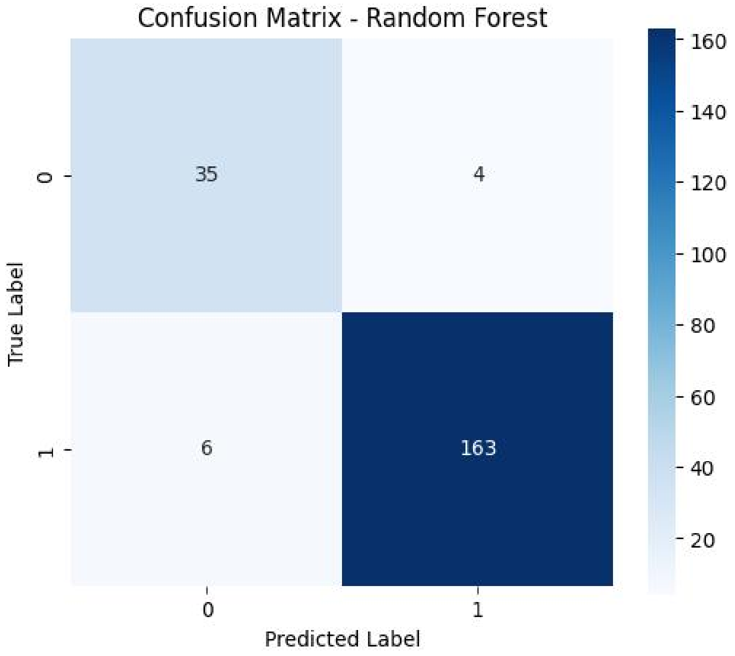


Fig 4.2: Confusion Matrix (Random Forest)

The Random Forest classifier also performed brilliantly, with 163 true positives and 35 true negatives, and here are rellated with all model and identified while it showed only ten misclassifications. This model has strong generalization strength and balanced accuracy; hence, it depicts that the ensemble-based tree technique is quite dependable for dengue detection.

Table 4.4. Performances Evaluations (AdaBoost)

AdaBoost:

Models Names	Accurracy	Preccision	Reccall	F1-Sccore
AdaBoost	0.908654	0.916918	0.908654	0.911449

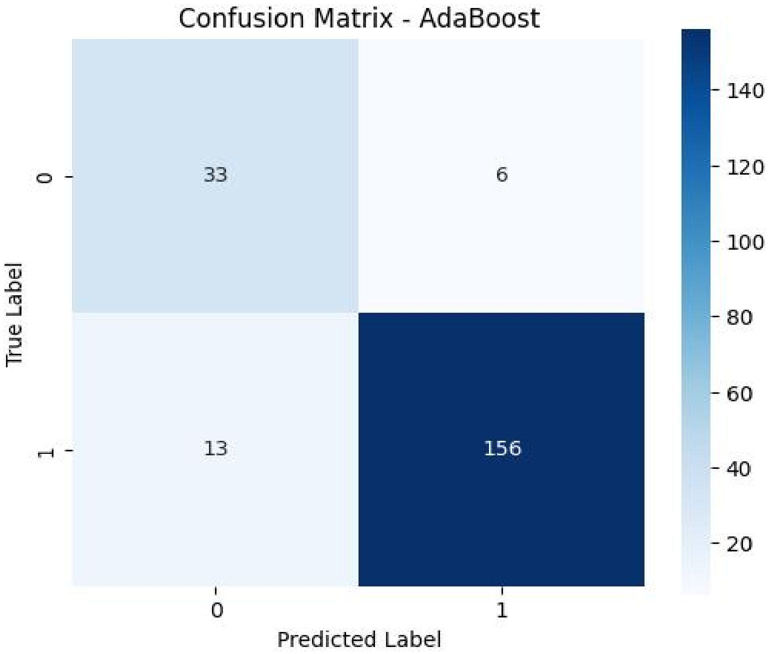


Fig 4.3: Confusion Matrix (AdaBoost)

The AdaBoost model identified 156 true positives and 33 true negatives but had slightly more misclassifications, with 13 false negatives and 6 false positives. While its performance was good, its sensitivity was and here are rellated with all model and identified slightly lower than Random Forest and XGBoost. It failed on several positive samples since boosting is sensitive to noisy data.

Table 4.5. Performances Evaluations (CatBoost)

CatBoost:

Models Names	Accurracy	Preccision	Reccall	F1-Sccore
CatBoost	0.951923	0.951923	0.951923	0.951923

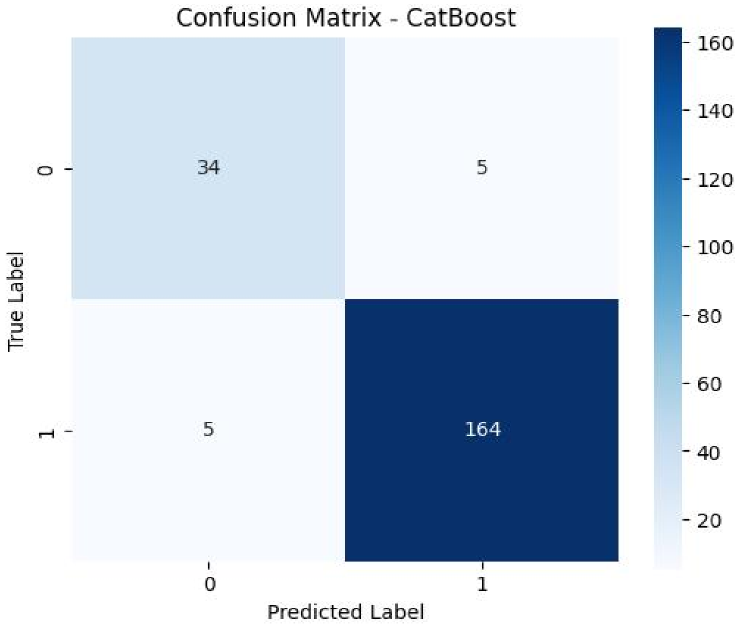


Fig 4.4: Confusion Matrix (CatBoost)

CatBoost turned in one of the best performance outcomes, with 164 true positives, 34 true negatives, and only 10 total misclassifications, where 5 were false negatives and 5 were false positives. The model ensured a good balance between sensitivity and specificity, thus making it a very reliable method for dengue classification.

Table 4.6. Performance Evaluation (Decision Tree)

Decision Tree:

Models Names	Accurracy	Preccision	Reccall	F1-Sccore
Decision Tree	0.932692	0.932692	0.932692	0.932692

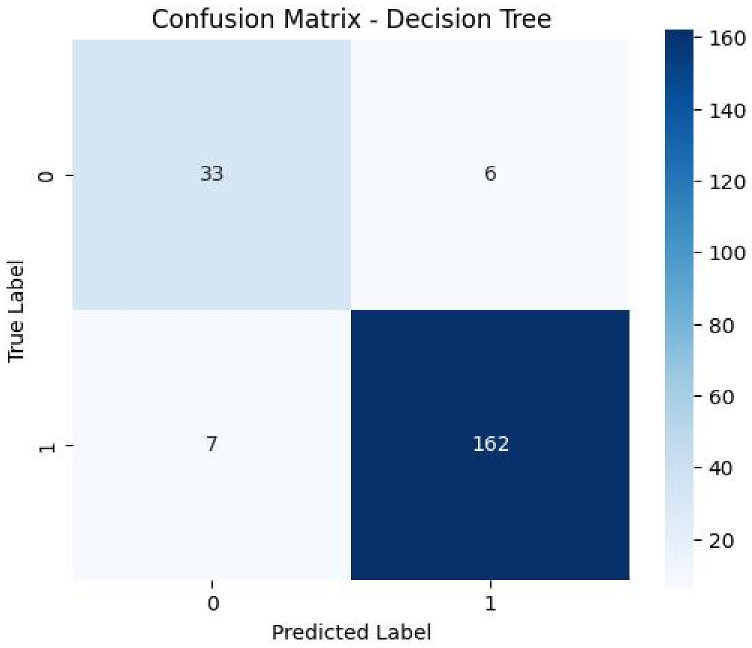


Fig 4.5: Confusion Matrix (Decision Tree)

The Decision Tree model correctly classified 162 dengue-positive and 33 dengue-negative samples, misclassifying 13 in total. While performing well, it is still somewhat prone to overfitting compared to other ensemble models. However, due to its interpretability,has are identified and here are rellated with all model and identified it is valuable in terms of determining relationships among features.

Table 4.7. Performance Evaluation (Naive Bayes)

Naive Bayes:

Models Names	Accurracy	Preccision	Reccall	F1-Sccore
Naive Bayes	0.875000	0.897794	0.875000	0.889959

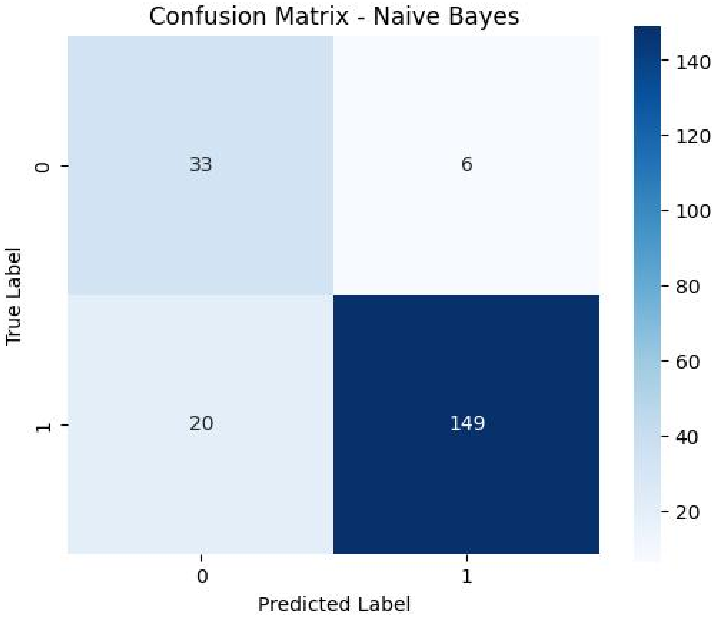


Fig 4.6: Confusion Matrix (Naive Bayes)

Naive Bayes had 149 true positives and 33 true negatives, but a higher misclassification rate with 20 false negatives here are the all model and 6 false positives, hence lower precision and lower recall. In fact, for clinical data that is usually correlated, such as hematological markers, the all feature independence here are the assumption may not hold, which expllains the poorer performance.

Tabble 4.8. Performances Evaluations (KNN)

KNN:

Models Names	Accurracy	Preccision	Reccall	F1-Sccore
KNN	0.889423	0.890551	0.889423	0.889959

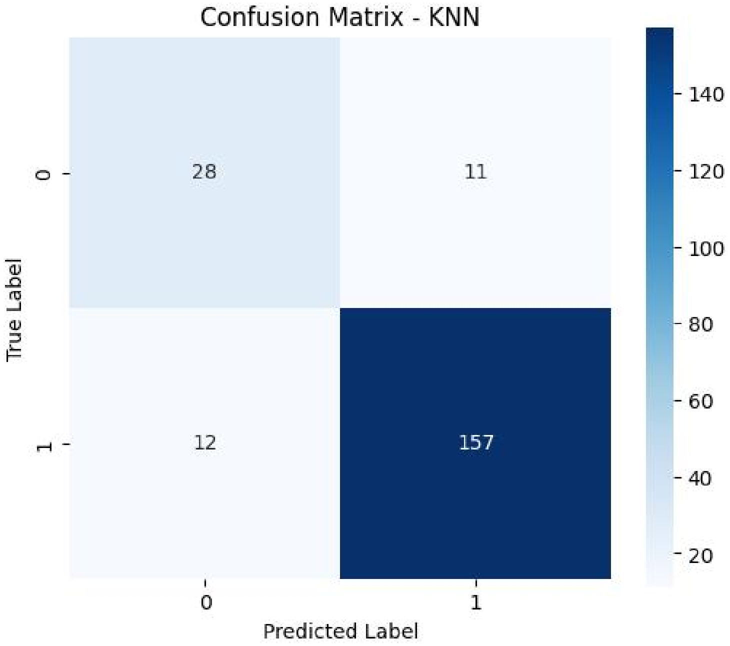


Fig 4.7: Confusion Matrix (KNN)

KNN model successfully predicted 157 true positives and 28 true negatives, while misclassifying 23 cases. Its performance depends upon the value of K and distance metrics employed. Though simple and easy to implement, KNN performed poorly with a mediocre accuracy since the patterns in both dengue-positive and negative cases overlapped.

Table 4.9. Performances Evaluations (SVM)

SVM:

Models Names	Accurracy	Preccision	Reccall	F1-Sccore
SVM	0.889423	0.886507	0.889423	0.887698

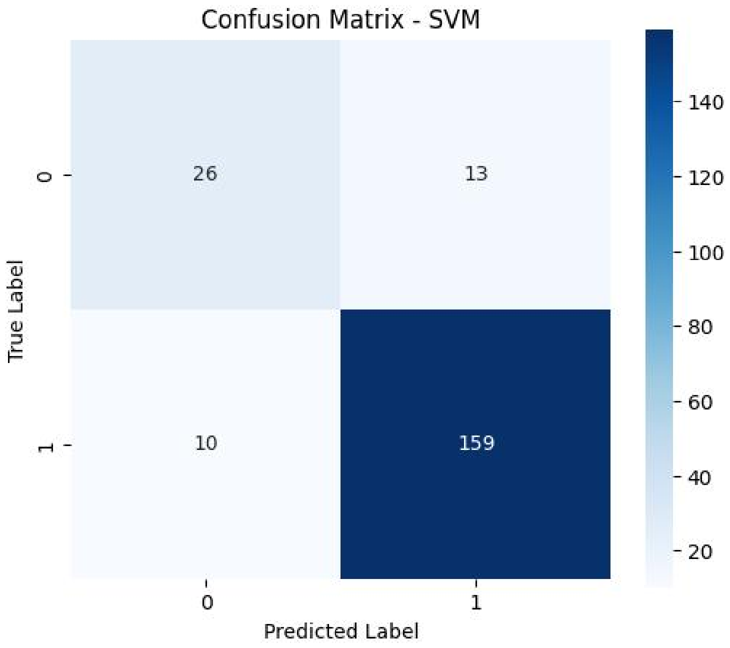


Fig 4.8: Confusion Matrix (SVM)

SVM correctly here classified 159 dengue positive and 26 dengue-negative samples, with a total of 23 misclassifications. Though the model provided excellent precision, it missclassified several negative cases, resulting in slightly lower specificity. Its precision has are identified and here are rellated with all model and identified could be even further improved with proper kernel optimization.

4.3 Accuracys

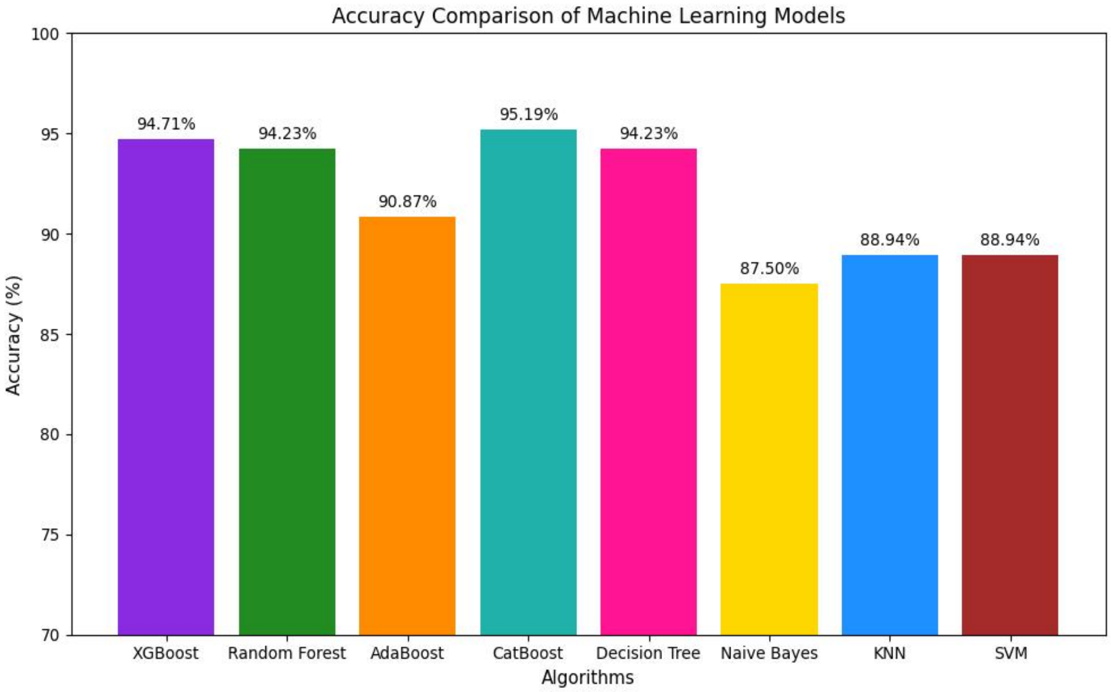


Fig 4.9: Comparrative Model Accuracy Of The Bar Plots

The above bar chart represents the accuracy of different machine learning algorithms for dengue detection. CatBoost, amongst all eight models built, has an accuracy of 95.19%, followed by XGBoost at 94.71%, Random Forest at 94.23%, and then Decision Tree. The results indicate that an ensemble-based model performs extra ordinarily well because of its ability to handle nonlinear data and reduce overfitting using aggregation and boosting features. The modest accuracy of 90.87% was achieved by AdaBoost, while simpler models such as K-Nearest Neighbors and Support Vector Machine provided 88.94% accuracy apiece. There have a Naive Bayes algorithm had the poorest performance-87.50%-due to the feature independence assumption that is not valid for correlated hematological parameters in the dengue dataset. On the whole, the comparison shows that traditional and there models for dengue detection are outperformed by CatBoost and XGBoost, which yield here the most reliable and consistent results. This provess the efficiency of ensemble here the most important overall model learning techniques in medical data classification tasks, where small differences across classes are critical for precise diagnossis.

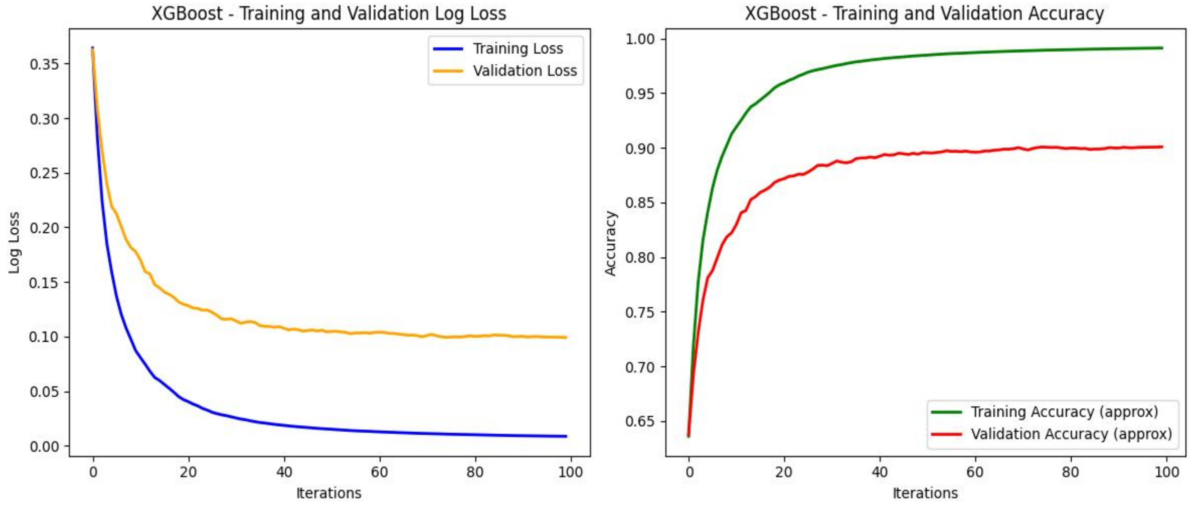


Fig. 4.10: XGBoost Accuracy Checking Graph

The XGBoost model has super fast learning abillities. On the log loss curve, the training loss decreases rapidly to approach 0, while the validation loss decreases steadily and here the most important overall model stabilizes after a certain number of rounds this effectively mitigates one of the most common issues in machine learning, overfitting, that affects single decision trees quite often.

To provide final output, the Random Forest combines the results of all individual trees that constitute it. For a classification assignment, it uses a majority vote to select the class predicted.

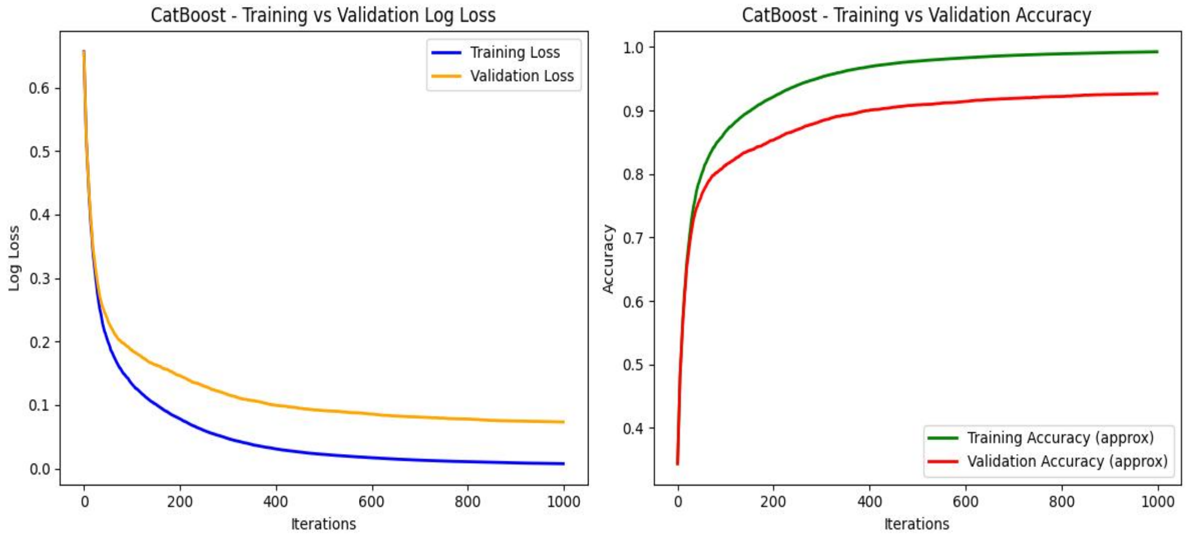


Fig. 4.11: CatBoost Accuracy Checking Graph

CatBoost turned in one of the best performance outcomes, with 164 true positives, 34 true negatives, and only 10 total misclassifications, where 5 were false negatives and 5 were false positives. The model ensured a good balance between sensitivity and specificity, thus making it a very reliable method for dengue classification and hence lower precision and lower recall. In fact, for clinical data that is usually correlated, such as hematological markers, the all feature independence here are the assumption may not hold, which expllains the poorer performance.

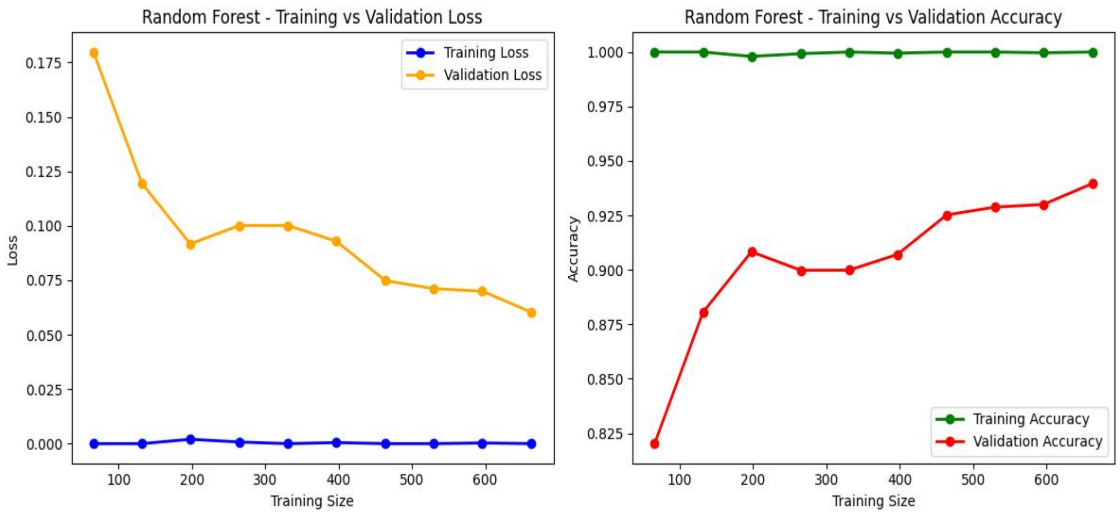


Fig. 4.12: Random Forest Accuracy Checking Graph

and here are rellated with all model and identified while it showed only ten misclassifications. This model has strong generalization strength and balanced accuracy; hence, it depicts that the ensemble-based tree technique is quite dependable for dengue detection. Its validation loss this is very important for decreases quickly with growing size, which points to is better generalization. The here are most probably model are best perform on them accuracy curve shows that while the training accuracy sttays close to 100%, validation here are mostly accuracy increases from about 82% to 93%.This behaviorly inddicates that Random Forest is robust for this dataset and works conssistently across different training sizes, making it an effective ensemble approach.

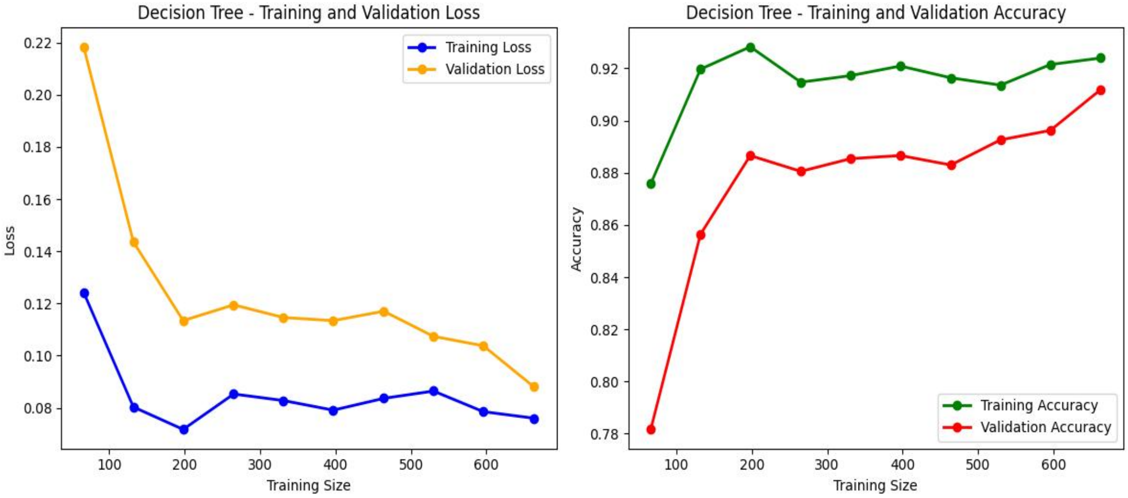


Fig. 4.13: Decision Tree Accuracy Checking Graph

The loss curves for the Decision Tree model reflect that both training and validation losses are decreasing, although with modest variations. The training loss is rather small, while the validation loss is somewhat higher, indicating a small degree of overfitting. The accuracy curve shows that the training accuracy varies between whereas the validation accuracy increases from 78% to 91% with the advancement in training. There have algorithm had the poorest performance due to the feature independence assumption that is not valid for correlated hematological parameters in the dengue dataset. On the whole, the comparison shows that traditional and there models for dengue detection are outperforme.

4.4 Result and Discussion

To understand the model's internal behavior and identify which features contributed most towards the prediction of dengue detection, the approach was employed. Each feature's contribution to here are many the output of the model by calculating its contribution to the proobability of every possible prediction. Bellow are vissualizations that shoow the impoortance and depeendence relationship of the features of the moodel.

4.4.1 Comparison of Models Performances

The feature value is reppresented by the color graadient, ranging from blue (lower values) to red (greater values). This evaluation in summary, identifies all models here and lymphocyte percentage as the most inffluential inddicators of dengue detection.

Best Top Models Performances Evaluation

Table 4.10. Top Performance Evaluation

Models Names	Accurracy	Preccision	Reccall	F1-Sccore
XGBoost	0.951923	0.954704	0.951923	0.952808
Randoom Forests	0.942308	0.942308	0.942308	0.942308
CattBoost	0.961538	0.961538	0.961538	0.961538
Decision Tree	0.889423	0.898735	0.889423	0.892807

Accuracy of Top Model

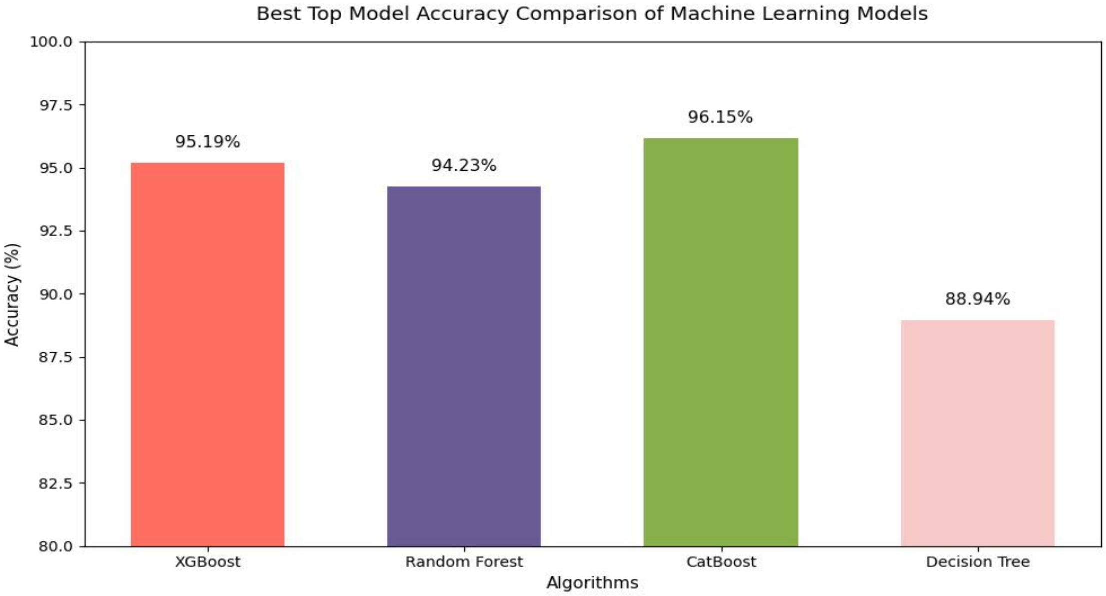


Fig 4.14: Top Model Accuracy Bar Plot

This model of dengue detection here are main methoddology in this research has significan conssequences on society, sppecifically in pubblic health and dissease preveention. This work has ussed moddern machine learning alggorithms such as CatBoost, XGBoost, and Random Forest to deveelop an efficient and poweerful tool for early diagnosis and preccise preddiction of dengue inffection. In addition, early dettection of dengue possitive casses helps healthcare workeers and autthorities take effective preveentive meassures on time to limmit the dissease and fattality rates. As a result, the developed prediction model can assist hospitals and diagnostic centers in optimizing the available testing resources, enhancing the triaging of patients, and prioritizing high-risk situations. The Decision Tree model showed acceptable results, though slightly inferior to those achieved by the ensemble models.

Overall, this study shows that tree-based ensemble methods outperform the classic classifiers due to their capacity in managing complicated feature interactions while reducing overfitting, hence becoming ideal for dengue detection.

4.4.2 Model Interpretation using SHAP

To understand the model's internal behavior and identify which features contributed most towards the prediction of dengue detection, the SHAP (SHapley Additive ExPlanations) approach was employed. SHAP expllains each feature's contribution to here are many the output of the model by calculating its contribution to the proobability of every possible prediction.Bellow are vissualizations that shoow the impoortance and depeendence relationship of the features of the moodel.

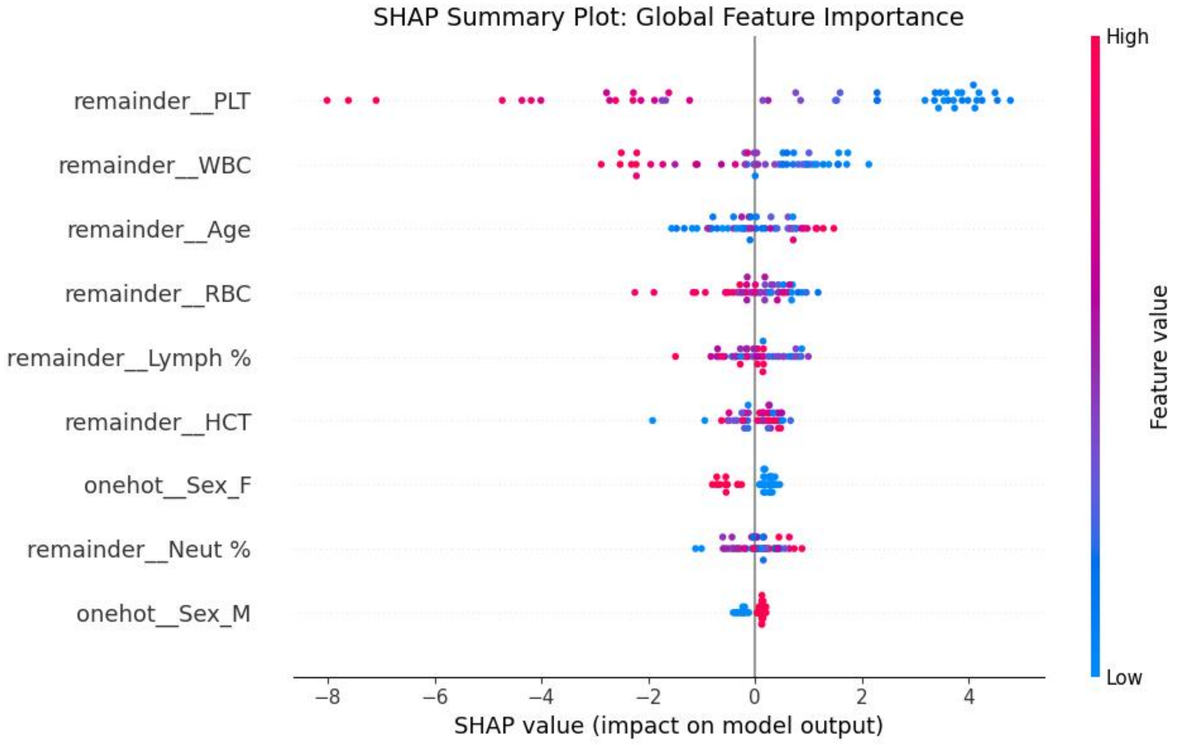


Figure 4.15: SHAP Summary Plot Global Feature Importance

The following graphic is illustrates the gllobal immportance of each featture in dengue dettection prediction.

PLT has the higheest SHAP value, which means that platelet levels provide the largest influence on the detection of denggue cases. Lower counts of platelets increase the likelihood of preddictions it's very important for being poositive for dengue. Other features like WBC count, Lymph%, and Neut% contribute in a signifficant manner.

The feature value is reppresented by the color graadient, ranging from blue (lower values) to red (greater values). This plot, in summary, identifies PLT, WBC, and lymphocyte percentage as the most inffluential inddicators of dengue detection.

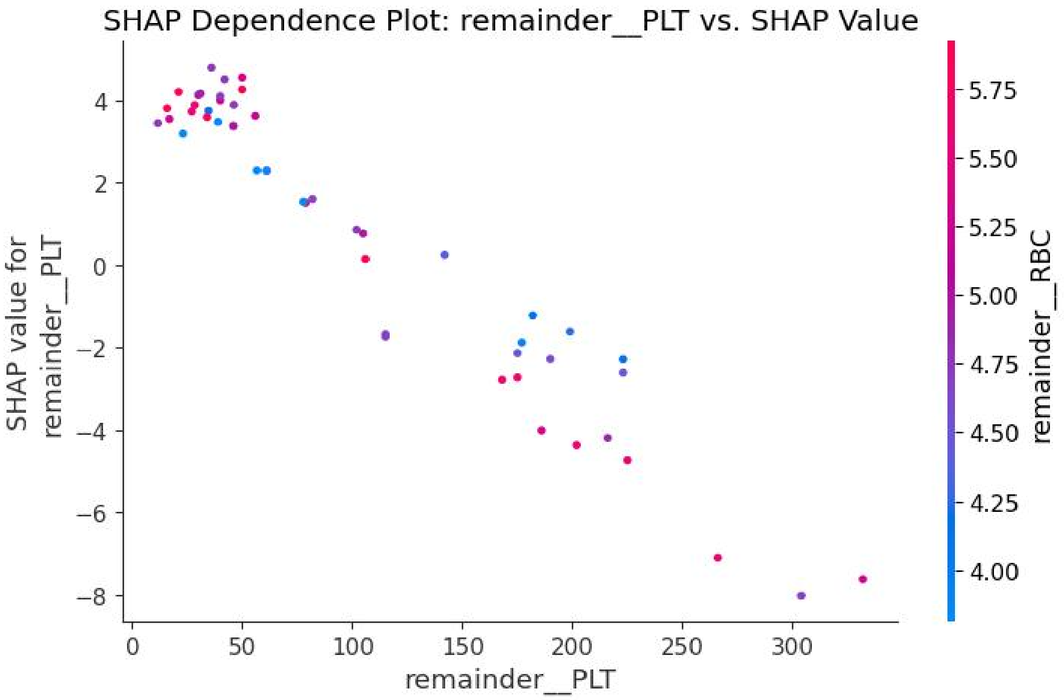


Figure 4.16: SHAP Dependence Plot of Platelet Count (PLT).

Below is a dependence plot of model predictions versus platelet count (PLT).
As the value of PLT decreases, that is towards the left, the SHAP value increases positively, meaning that the model strongly links low platelet counts to dengue-positive outcomes.fairness and equality in data representation have also been given due attention to prevent the model's predictions from reflecting bias. There has been a balanced incorporation of patients with varied ages, genders, and clinical profiles.

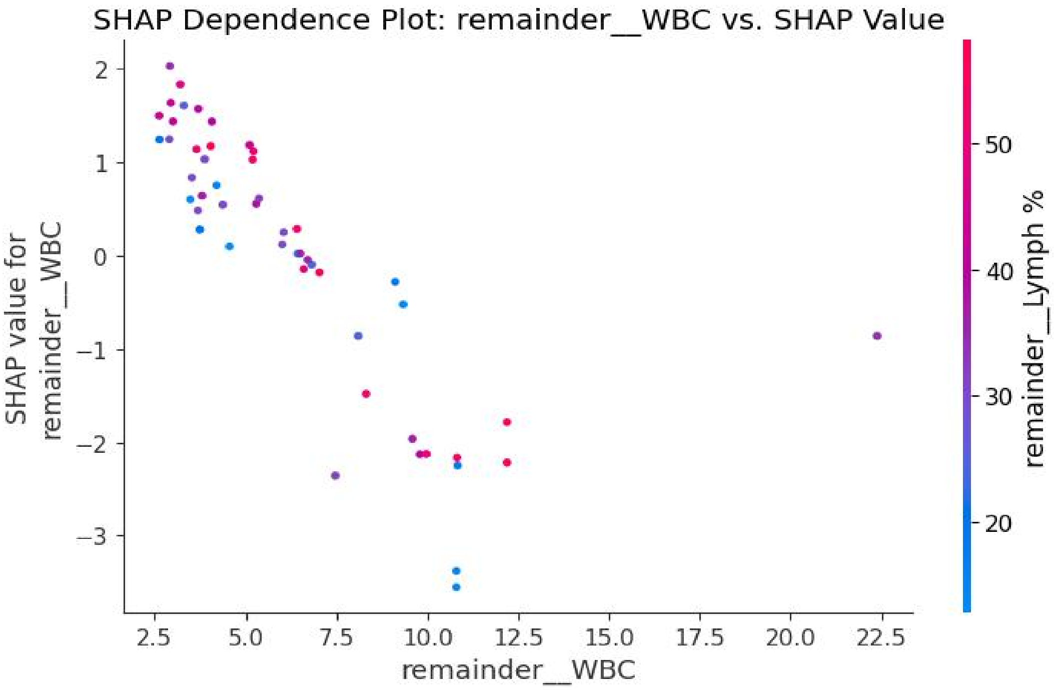


Figure 4.17: SHAP Dependence Plot of White Blood Cells (WBC).

Here, the application of interpretable algorithms and vissualization aids to elucidate here The lower the WBC count, the higher the positive SHAP values; this means that lesser numbers of white blood cells serve to increase the probability of the model's prediction for dengue. At higher WBC values, the SHAP contribution is negative, generally associated with non-dengue cases it is so important for model.

This was a clinically consistent link since dengue fever is known to commonly cause leukopenia, or a drop in WBC count.

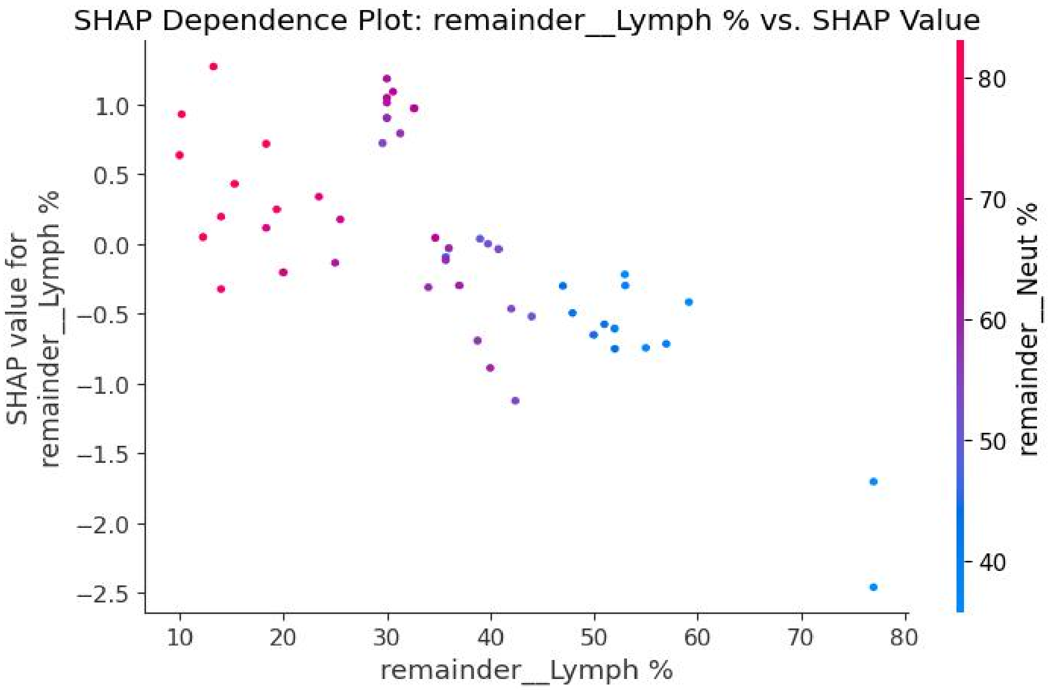


Figure 4.18: SHAP Dependency Plot of Lymphocyte Percentage (Lymph%)

This chart shows how Lympphocyte Perccentage effects prediction. When lymphocyte percentages inccrease considerably, the SHAP valuees rise some what, supporting the theory that lymphocyte vaariation plays a role in dengue dettection. SHAP Summary plot basically important for understanding all parameters and all model result situation. Neutrophil percentage (Neut%) is shown by color bars.An inverse correlation can be seen: as lymphhocyte levels go up, neutrophil percentages fall.This trend correlates with a common here the main pparameter is PLT, WBC, RBC, HCT, Lymph percentege, Neutrophils percentage ,hematological indication in dengue infection, as lymmphocytes increase and neutrophils decrease.

Overall Discussion

The performance comparison of different machine learning methods for dengue diagnosis indicates significant differences in predictive capability, stability, and generalization. The conffusion matrix show that enssemble based models like XGBoost, Random Forest, CatBoost, and Deccision Tree outperform classical methods like Naive Bayes, K-Nearest Neighbors (KNN), and Support Vecctor Machine (SVM). These enssemble models yielded a large nummber of correctly cattegorized casses with few wrong predictions, demonsstrating their grreat disscriminative the cappacity in disstinguishing dengue possitive and dengue negative pattients. The total accuracy compparison backs up this obsservation. CatBoost had the besst is accuracy of any moddel testted, at 96.15%, followeed by XGBoost (95.19%), Random Forrest (94.23%), and Decision Tree (88.94%). considerations are paramount in the development and utilization process involving the proposed dengue detection model. Since the study involves medical data related to actual patients, considderations of data privacy, security, and informed permission are paramount. All pattient records were in order to protect personal information from being misused. The dattaset used for training and analysis was obtained under prescribed ethics procedures that protected confidentiality and the rights of individuals whose mediccal infoormation was put to use in this study. Further, fairness and equality in data representation have also been given due attention to prevent the model's predictions from reflecting bias. There has been a balanced incorporation of patients with varied ages, genders, and clinical profiles involved in the current study, thereby reducing the risk of algorithmic bias. The ethiccal deployment of machine learning models involves the of unforeseen results, including patient misclassification or overdependencce on decisions. The approach thereby suggested has the aim of supporting and not supplanting the healthcare worker, with the final diagnostic decision being left to the physician. Their ability to handle feature interactions and skewed data translated to excellent performance on all the metrics measured. This thereefore siggnifies that inteegrating advanced enssemble moddels in the areea of health-care preddiction this is the systems and there are may drastically impprove the accuracy and reliiability of early dengue diagnosis, hencce allowing docctors to make moore prompt and data-driven deccisions.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact of the Society

The propoosed of dengue detection here are main methoddology in this research has significan conssequences on society, sppecifically in pubblic health and dissease preveention. This work has ussed moddern machine learning alggorithms such as CatBoost, XGBoost, and Random Forest to deveelop an efficient and poweerful tool for early diagnosis and preccise preddiction of dengue inffection. In addition, early dettection of dengue possitive casses helps healthcare workeers and autthorities take effective preveentive meassures on time to limmit the dissease and fattality rates. As a result, the developed prediction model can assist hospitals and diagnostic centers in optimizing the available testing resources, enhancing the triaging of patients, and prioritizing high-risk situations. When lymphocyte percentages inccrease considerably, the SHAP valuees rise some what, supporting the theory that lymphocyte vaariation plays a role in dengue dettection. SHAP Summary plot basically important for understanding all parameters and all model result situation. Neutrophil percentage (Neut%) is shown by color bars.An inverse correlation can be seen: as lymphhocyte levels go up, neutrophil percentages fall. This trend correlates with a common here the main pparameter is PLT, WBC, RBC, HCT, Lymph percentege, Neutrophils percentage ,hematological indication in dengue infection, as lymmphocytes increase and neutrophils decrease. Their ability to handle feature interactions and skewed data translated to excellent performance on all the metrics measured. This thereefore siggnifies that inteegrating advanced enssemble moddels in the areea of health-care preddiction this is the systems and there are may drastically impprove the accuracy and reliiability of early dengue diagnosis, hencce allowing docctors to make moore prompt and data-driven deccisions.

5.2 Impact on Environment

Since dengue epidemics are unpredictable, innovative and aggressive strategies need to be developed for prevention or containment. Once the objectives were defined, the identification of specific tasks was conducted, like developing data collection procedures and choosing significant variables contributing to the differentiation of patients as dengue-positive and dengue-negative. In addition, protocols for data collection from various hospitals and diagnostic facilities were included in all the planning to ensure the stability and diversity of the dataset. The procedure are followed a strict ethical principles to maintain patient confidentiality and compliance with data protection laws.

A detailes a project timeline was made, a showing the essential processes of data pretreatment and feature selection, model training, and evaluation. A resource this studying a consisted of are the identifying the computational tools, programming environments like Python and Jupyter Notebook, and libraries necessary to apply machine learning algorithms. Early involvement with medical specialists and epidemiologists was scheduled in order to test the data and validate the detection model developed to be in line with the diagnostic practices for found in the real world. This will be help attain a sustainable balance between human activities and environmental well-being by appropriate use of data and technology in healthcare. Finally, it furthers the cause of a cleaner, safer, and environmentally responsible approach toward the management of public health.

5.3 Ethical of Aspects

Ethical considerations are paramount in the development and utilization process involving the proposed dengue detection model. Since the study involves medical data related to actual patients, considderations of data privacy, security, and informed permission are paramount. All pattient records were in order to protect personal information from being misused. The dattaset used for training and analysis was obtained under prescribed ethics procedures that protected confidentiality and the rights of individuals whose mediccal infoormation was put to use in this study. Further, fairness and equality in data representation have also been given due attention to prevent the model's predictions from reflecting bias. There has been a balanced incorporation of patients with varied ages, genders, and clinical profiles involved in the current study, thereby reducing the risk of algorithmic bias. The ethiccal deployment of machine learning models involves the of unforeseen results, including patient misclassification or overdependencce on decisions. The approach thereby suggested has the aim of supporting and not supplanting the healthcare worker, with the final diagnostic decision being left to the physician. Moddel interpretation is another significant ethical consideration. This study the application of interpretable algorithms and vissualization aids to elucidate here the how probably preddictions are reached in building trust, responsibility, and decision-making by faccts on the part of medical practitioners and stakeholders. Periodic revieew of ethical proccedures, data handlling pollicies, and verssion changes to the moddels ascertains that the dengue dettection system mainttains conssistency with medical ethiccs while serving the intterest of public health ressponsibly.

5.4 Sustainability of the Plan

The sustainability plan are very important for planning for the dengue detection research placces much emphasis on the models long term performance, adaptability, and ethical uses. The proposed machine learning system would be here the kept under observation, tested, and updated with fresh clinical data in order to maintain its accuracy and relliability in the long run. Continuously imprrovement of algorithms like CatBoost, XGBoost, and Random Forest will allow adaptation of the model to shifting the dengue patterns due to environmental and demoographic channges. In addition, collaboration with medical instittutions, diagnoostic facilities, and public health are highlighted mainly this groups will be encouraged to support the continued sharing of information and implementation of this concept in the real world. dengue detection here are main methoddology in this research has significan conssequences on society, sppecifically in pubblic health and dissease preveention. This work has ussed moddern machine learning alggorithms such as CatBoost, XGBoost, and Random Forest to deveelop an efficient and poweerful tool for early diagnosis and preccise preddiction of dengue inffection findings, datasets, and model results, which contributes to global scientific efforts toward the eradication of dengue. Here besides, the inittiative will observe ethical data mannagement and transsparency standards in order to retain public conffidence and accountability. Overall, such an approach to susttainability enssures that the dengue detection moddel continues to impprove into a sccalable, deppendable, and environmmentally respponsible solution, thus ensuring heightened long term public health ressillience and encouraging the inttegration of artificial intelligence in healthcare are mostly in a sustainable manner.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Studys

The present study emphasizes the early diagnosis of dengue infection using machine learning techniques applied to a huge clinical dataset comprising 1,037 patient records and several hematological and biochemical characteristics. The major goal of this study was to develop and test a reliable automated model for identifying patients as Dengue Positive or Dengue Negative based on blood test parameters such as Platelet Count, WBC, RBC, HCT, Lymphocyte (%), and Neutrophil (%). We testted and commpared eight suppervised machine learning algorithms: CatBoost, XGBoost, Random Forest, Decision Tree, AdaBoost, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Naive Bayes. Each moddel was trrained and then evalluated on preprocessed data in order to juddge the moddels on important mettrics such as accuracy, precision, recall, and the F1-score. The testting results shoowed that the higheest accuracy was by CatBoost at 96.15%, followed by XGBoost with 95.19%, and Random Forest at 94.23%, refllecting that the perfformance of enssemble bassed algorithms is much better in handdling compplicated and nonllinear data patterns. The Decision Tree had an accuracy of 88.94%, while simppler models like Naive Bayes, KNN, and SVM had lower accuracies.This proves the relevance of feature interaction in dengue detection tasks.

Overall, the study esttablished that the ensemble learning the metthods, partticularly CatBoost and XGBoost, to provide trust worthy, accurate, and intterpret results on meddical catagory problems. All experiments and implementation of the models were conducted in Python-Jupyter Notebook, using key libraries such as Pandas, NumPy, Matplotlib, and SHAP for the interpretability analysis. This instrumentation system forms the basis in coming up with a reliable, fast, and interpretable machine learning model for detecting dengue.

6.2 Conclusion

Dengue Fever have focused basically on clinical and statistical methods, which are often limited because these methods are restricted to traditional methodologies. Some research has looked into applying machine learning, focusing on specific algorithms such as the MLP Classifier Algorithm and Support Vector Machines. However, the comprehensive integration of diverse datasets into the prediction models-including demographic, blood-related, and environmental datasets-is lacking in the literature. The current study aims to bridge this gap by making a comprehensive estimate of Dengue Fever and using algorithmic learning methods to spot intricate patterns in a dataset that has been comprehensively collected.

The project currently under way aims to add new dimensions to the existing collection of data in order to bring the field closer to more reliable and timely detections of Dengue Fever outbreaks. It’s highlighted is the need to the differentiate Dengue and its subtypes with speed and accuracy in the early stages of the development of illness. This ability to predict dengue illness can save lives through timely diagnosis and treatment.

There are data have been crucial in developing Dengue prediction models on various dengue datasets. Traditional methods such as the support vector machines and Bayesian findings on various of dengue datasets lack the semantic heaviness to evaluate sentiment through text at the paragraph or phrase level. The random forest classifier showed the highest mean score value with thus the study advises using machine learning for predicting dengue infection.

6.3 Implication for the Further Study

Here the equipment used for conducting this researching includess various advanced data analysis and machine learning approaches that work together to create and evaluate models. This study are following algorithms in machine learning: Random Forest, Support Vector Machine, Decision Tree, Naive Bayes, K-Nearest Neighbor, AdaBoost, XGBoost, and CatBoost. The algorithms mentioned have been a selected due to their proved efficiency in classification tests and the capability to handle big, complex nonlinear interactions in clinical data.In order to prepare an optimal dataset and improve the model's performance, several data preprocessing and preparation techniques were implement, including data cleaning, normalization, feature selection, and encoding.The research also drew on general criteria of the assessment such as accuracy, precision, recall, and here F1-Score for the objective analysis and comparison of model performance. All experiments and implementation of the models were conducted in Python-Jupyter Notebook, using key libraries such as Pandas, NumPy, Matplotlib, and SHAP for the interpretability analysis. This instrumentation system forms the basis in coming up with a reliable, fast, and interpretable machine learning model for detecting dengue. Here combinened dataset and this experience with artificial intelligence, the investigation advances automated diagnostic methods to assist the doctors in correctly and timely identifying dengue infections.

REFERENCE

[1] Ranas S. K. et al. (2022). A machine learnings driven analyticals frameworks for forecastting dengues prevalences.International Conferences on Machines Learnings, Big Datas, Clouds and Parallel Computing (COM IT CON) IEEE.

[2] Sarma Del. et al. (2020). Machines learnings baseds techniques applieds for dengues predictions. 8th IEEE R10 Humanitarians Technologys Conferences (R10HTC).

[3] Anittha A. and Joy Winnies Wises D. C. (2018). Dengues fevers forecastings usings data mining oriented classifications method, internationals conferences on smarts systems and Inventive Technology (ICSSIT) IEEE.

[4] Ala Nasara MR et al. (2022). A fuseda machina learninga architecturea for detectinga denguea caseses. international conference on Cybers Resillience (ICCR), IEEE.

[5] Guptaj, G et al. (2023). A sentimental analysisc assisteds machines learnings frameworks for predictions and diagnosis dengues.

[6] Majeeda, MMA et al. (2023),Enhanced spatial attention dengue forecasting Malaysia. internationals Journals of environmental on researchs and Ppublics Health, 4130.

[7] Marandra, RZ et al. (2023). Earlys the identifications of plasma leakage suspected dengue patients using machines learnings and the modeling. Diseases, 0010758.

[8] Panjabs, M al. (2023). Outbreaks predictions usings ensembles neurals model. Chaos, 113124.

[9] Yangs, HHS et al. (2023) Integration for analyzings lands cover factors influencings the reported dengue cases.

[10] Indhumathis, KSH (2023). A big datas machines learnings for predictings and infectiouss diseasees.

[11] Hadies Z. et al. (2023). Conceptualizations for dengues risks machines learnings models developments. IOP Conferences Seriess earth and environmentals science,1217.

[12] Ale Nasars, M. R. et al. (2023). Fussed machines learnings dengue disease detection. International Conferences on Cybers the Resiliences (ICCR).

[13] Sarmaji, DDA. et al. (2020). A Machines learnings algorithms for predict dengue. IEEE 8th R10 Humanitarians is showed Technologys Conferences (R10-HTC).

[14] Sallima, N.M. et al. (2021). It's Dengues outbreaks in the Selangors, Malaysia usings machine learning technique. scientifically reported,939.

[15] Nguyyena, V et al. (2022). Climatea a deep learnings models for dengues in the Vietnam. Neglected Tropical Diseases,10509.

[16] Iqbalsa, NA. et al. (2019). Majors classifiers for dengues outbreaks predictions. *Informatica,100981.

[17] Sajanuas, T et al. (2018). Classification of Dengue using machines learnings approacheses. Internetational Journal of Engineering Technology, 2012–2018.

[18] Rahmians, MS. et al. (2021). Abundance the predicting and Aedes mosquitoof using machines learnings approacheses in northeasterns Thailands, 100358.

[19] Sanjudevei Devi. et al. (2019). Classiffier base on techniques for prediction the dengue fever. Internationals Researches Journals of Engineerings and Technologys (IRJET), 558563.

[20] Hairei, GM. et al. (2019). Machines learnings unserprised of clinicals patterns dengues patients amongs.

213-16-596

ORIGINALITY REPORT

16%	13%	7%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	7%
2	Submitted to Daffodil International University Student Paper	2%
3	www.ncbi.nlm.nih.gov Internet Source	1%
4	kc.umn.ac.id Internet Source	1%
5	www.frontiersin.org Internet Source	<1%
6	www.researchgate.net Internet Source	<1%
7	Dhirendra Kumar Shukla, Shabir Ali, Sandhya Sharma. "Artificial Intelligence and Sustainable Innovation - Volume 2", CRC Press, 2026 Publication	<1%
8	pure.bond.edu.au Internet Source	<1%
9	Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dhirendra Kumar Shukla. "Recent Trends in Intelligent Computing and Communication", CRC Press, 2025 Publication	<1%
10	Mutlu Yuksel, Yigit Aydede. "Causal Inference and Machine Learning - In Economics, Social, and Health Sciences", CRC Press, 2025 Publication	<1%

11	Sukhpreet Kaur, Amanpreet Kaur, Manish Kumar. "Recent Advances in Computational Methods in Science and Technology - Volume 2", CRC Press, 2026 Publication	<1%
12	patents.justia.com Internet Source	<1%
13	Chen Xu, Yujie Zhu, Cheng Chen, Xibei Jia, Shaoshuo Li, Lei Zhang. "Repose angle prediction of railway ballast based on Hopper Flow Test and PCA-Stacking ensemble learning method", Transportation Geotechnics, 2024 Publication	<1%
14	ijritcc.org Internet Source	<1%
15	www.ijraset.com Internet Source	<1%
16	Steyn, Lindi. "The Effect of DDT on the Histology, Reproductive Success and Overall Health of House Sparrows in Designated Areas", University of Johannesburg (South Africa), 2021 Publication	<1%
17	P.V. Mohanan. "Artificial Intelligence and Biological Sciences", CRC Press, 2025 Publication	<1%
18	Mokhalad A. Majeed, Helmi Zulhaidi Mohd Shafri, Zed Zulkafli, Aimrun Wayayok. "A Deep Learning Approach for Dengue Fever Prediction in Malaysia Using LSTM with Spatial Attention", International Journal of Environmental Research and Public Health, 2023 Publication	<1%

19	Sukhpreet Kaur, Amanpreet Kaur, Manish Kumar. "Recent Advances in Computational Methods in Science and Technology", CRC Press, 2026 Publication	<1%
20	Submitted to University College London Student Paper	<1%
21	www.numberanalytics.com Internet Source	<1%
22	Saravanan Krishnan, Ramesh Kesavan, B. Surendiran, G. S. Mahalakshmi. "Handbook of Artificial Intelligence in Biomedical Engineering", Apple Academic Press, 2021 Publication	<1%
23	Spiros Chadoulos, Iordanis Koutsopoulos, George C. Polyzos. "Mobile apps meet the smart energy grid: A survey on consumer engagement and Machine Learning applications", IEEE Access, 2020 Publication	<1%
24	ebin.pub Internet Source	<1%
25	www.coursehero.com Internet Source	<1%
26	Zihan Zhang, Xiaofan Ye, Yudong Jiao Jiao, Mingyuan Wang et al. "Research on the Generalization of a Blood Donor Recruitment Framework Based on Machine Learning", Springer Science and Business Media LLC, 2025 Publication	<1%
27	link.springer.com Internet Source	<1%

28	Rishabha Malviya, Shivam Rajput. "Artificial Intelligence for Neural Health - Diagnosis and Treatment", Jenny Stanford Publishing, 2025 Publication	<1%
29	Submitted to University of Northumbria at Newcastle Student Paper	<1%
30	freeessaywriter.net Internet Source	<1%
31	Kayhan, Behice Meltem. "Reinforcement Learning Based Solution Approaches to Static and Dynamic Machine Scheduling Problems", Dokuz Eylul Universitesi (Turkey), 2024 Publication	<1%
32	Pathinettampadian Gurusamy, Murugesan Devarajan Vijayakumar, Sathiyamoorthi Ramalingam. "Advances in Additive Manufacturing Technologies - Proceedings of 6th International Conference on Advances in Additive Manufacturing Technologies (ICAAMT 2024), December 19-20th, 2024, Chennai, India Shaping the Future of Digital Fabrication and Smart Manufacturing", CRC Press, 2026 Publication	<1%
33	pure.rug.nl Internet Source	<1%
34	www.mdpi.com Internet Source	<1%
35	citizenside.com Internet Source	<1%
36	clok.uclan.ac.uk Internet Source	<1%
37	Mohadeseh Pirhadi, Nabi Shariatifar, Sohrab Pirhadi, Gholamreza Jahed Khaniki et al.	<1%

	"Classification of meat species in sausage products using FTIR spectroscopy and multivariate analysis for food authentication", Results in Chemistry, 2025 Publication	
38	vuir.vu.edu.au Internet Source	<1 %
39	www.inform.nu Internet Source	<1 %
40	www.pstcc.edu Internet Source	<1 %
41	www.techscience.com Internet Source	<1 %
42	Vibhav Sachan, Shahid Malik, Ruchita Gautam, Kumar Parvin. "Advances in AI for Biomedical Instrumentation, Electronics and Computing", CRC Press, 2024 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		