

Machine Learning Approach to Predict Dengue Fever in Bangladesh

By

Morsalin Sarker Siam

213-15-4371

Md. Rashedul Islam

213-15-4308

FINAL YEAR DESIGN PROJECT REPORT

**This Report Presented in Partial Fulfillment of the Requirements
for the Degree of Bachelor of Science in Computer Science and
Engineering**

Supervised by

Dr. Arif Mahmud

Associate Professor and Associate Head

**Department of Computer Science and
Engineering Daffodil International University**

Co-Supervised by

Professor Dr. Sheak Rashed Haider Noori

Professor & Head

**Department of Computer Science and
Engineering Daffodil International University**



**DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh**

September 16, 2025

APPROVAL

This Project titled Machine Learning Approach to Predict Dengue Fever in Bangladesh submitted by Morsalin Sarker Siam (213-15-4371) and Md. Rashedul Islam (213-15-4308) to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 16-09-2025.

BOARD OF EXAMINERS



Professor Dr. Sheak Rashed Haider Noori (SRH)
Professor & Head
Chairman, Department of CSE, DIU

Board Chairman



Md. Sadekur Rahman (SR)
Assistant Professor
Department of CSE, DIU

Internal Member 1



Mr. Saiful Islam (SI)
Assistant Professor
Department of CSE, DIU

Internal Member 2



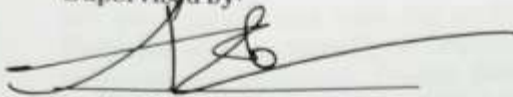
Dr. Md. Arshad Ali (DAA)
Professor
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science & Technology University

External Examiner

DECLARATION

We hereby declare that this project has been done by us under the supervision of Dr. Arif Mahmud, Associate Professor and Associate Head, Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:



Dr. Arif Mahmud

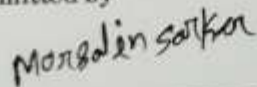
Associate Professor and Associate Head
Department of Computer Science and
Engineering Daffodil International University

Co-Supervised by:

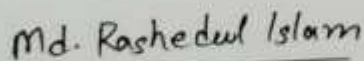


Professor Dr. Sheak Rashed Haider Noori
Professor & Head
Department of Computer Science and
Engineering Daffodil International University

Submitted by:



Morsalin Sarker Siam
ID: 213-15-4371
Department of Computer Science and
Engineering Daffodil International University



Md. Rashedul Islam
ID: 213-15-4308
Department of Computer Science and
Engineering Daffodil International University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has assisted us in one way or another.

First, we express our heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for us to complete the **Final Year Design Project (FYDP)** successfully.

We are grateful and wish our profound indebtedness to **Dr. Arif Mahmud, Associate Professor and Associate Head**, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor in the field of **Machine Learning** carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartfelt gratitude to the Head of the Department of Computer Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

We would like to thank our entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Fever and has been reporting outbreaks of the disease, it still continues to remain an important public health concern of the new millennium with significant morbidity strains on the healthcare system. It is important for early detection and error-free prevention of the disease. The included subjects comprised a group of patients admitted to multiple various healthcare centers linked with medical records (i.e. clinical patient data (gender, age), hematological data and other conventional clinical indicators). The dataset had been pre-processed to treat missing values and outliers in data provided, scaled for model learning. The models that are used in this study is Random Forest, Logistic Regression, Decision Tree, Support Vector Machines (SVM), Neighbors (KNN) K-Nearest, GaussianNB, XGBoost Stacking Ensemble. The study results indicated that Random Forest achieved the highest accuracy (0.89), followed closely by XGBoost (0.87) and the Stacking Ensemble (0.88). Decision Tree and Support Vector Machines (SVM) demonstrated moderate performance with accuracies of 0.82 and 0.83, respectively. Logistic Regression performed fairly well at 0.79, while comparatively lower accuracies were observed for K-Nearest Neighbors (KNN) at 0.76 and Gaussian Naïve Bayes (GaussianNB) at 0.74. The models were evaluated based on precision, recall, F1 score and confusion matrices where Random Forest and Logistic Regression performed the best in terms of minimizing false positives and false negatives. Finally, we have showed being useful machine learning models, in this specific Random Forest and Logistic Regression in the prediction of outbreaks of dengue. However, the results of this analysis suggest that the use of such models maybe beneficial for more efficient monitoring of surveillance, enhancing resource allocation and thus more timely interventions in controlling Dengue Fever in Bangladesh. Higher quality of real-time model parameter estimations with more types of input information may improve the reliability and support to decision makers.

Table of Contents

Approval	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
List of Figures	vii
List of Tables	viii
1 Introduction	1
1.1 Introduction	1
1.2 Motivation	1,2
1.3 Objectives	2
1.4 Methodology	2,3
1.5 Project Outcome	3
1.6 Organization of the Report.....	3,4
2 Background	5
2.1 Introduction	5
2.2 Literature Review.....	5-8
2.2.1 Similar Applications.....	9
2.3 Gap Analysis.....	9
2.4 Summary	10
3 Research Methodology	11
3.1 Methodology/Requirement Analysis & Design Specification	11,12
3.1.1 Overview	12,13
3.1.2 Proposed Methodology/ System Design.....	13
3.1.3 Functional and Nonfunctional Requirements	14
3.1.4 Data Flow Diagram.....	15
3.1.5 UI Design.....	16,17
3.2 Detailed Methodology and Design.....	18
3.3 Project Plan.....	18,19
3.4 Task Allocation.....	20
3.5 Summary.....	20
4 Implementation and Results	21
4.1 Environment Setup.....	21

4.2	Testing and Evaluation/Performance/ Comparative Analysis	21
4.3	Results and Discussion.....	22-34
4.4	Summary	35
5	Engineering Standards and Design Challenges	36
5.1	Compliance with the Standards	36
5.1.1	Software Standards	36
5.1.2	Hardware Standards	37
5.1.3	Communication Standards	37
5.2	Impact on Society, Environment and Sustainability	38
5.2.1	Impact on Life	38
5.2.2	Impact on Society & Environment	38
5.2.3	Ethical Aspects.....	39
5.2.4	Sustainability Plan	40
5.3	Project Management and Financial Analysis	40
5.4	Complex Engineering Problem	41
5.4.1	Complex Problem Solving	41
5.4.2	Engineering Activities	42
5.5	Summary	42
6	Conclusion	43
6.1	Summary.....	43
6.2	Limitation.....	43
6.3	Future Work.....	44
	References	45,46

List of Figures

3.1	Proposed Methodology.....	13
3.2	Data Flow Diagram.....	15
3.3.1	UI Design.....	16
3.2.2	Confusion Matrix for Logistic Regression.....	25
3.2.3	Confusion Matrix for Random Forest.....	26
3.2.4	Confusion Matrix for Decision Tree.....	27
3.2.5	Confusion Matrix for KNN	27
3.2.6	Confusion Matrix for SVC.....	28
3.2.7	Confusion Matrix for Gaussian Naive Bayes.....	29
3.2.8	Age.....	29
3.2.9	Q-Q Plots.....	30
3.2.10	Hemoglobin Levels.....	31
3.2.11	Pairplot of Age, Hemoglobin and Total WBC Count.....	31
3.2.12	Distribution of Hemoglobin.....	32
3.2.13	Distribution of WBC.....	32
3.2.14	Distribution of Hemoglobin of violin.....	33
3.2.15	Model Accuracy Comparison.....	34

List of Tables

2.1 Summary of Literature Reviewed.....	7,8
2.2 Gap Analysis.....	9
3.3 Project Plan.....	19
3.4 Task Allocation.....	19,20
4.2 Testing and Evaluation/Performance/ Comparative Analysis.....	21
4.2.1 Accuracy for Classification.....	22
4.2.2 Precision, Recall, F1-Score, and Support—Logistic Regression.....	23
4.2.3 Precision, Recall, F1-Score, and Support—Random Forest.....	23
4.2.4 Precision, Recall, F1-Score, and Support—Decision Tree.....	23
4.2.5 Precision, Recall, F1-Score, and Support— K-Nearest Neighbors (KNN).....	24
4.2.6 Precision, Recall, F1-Score, and Support—Support Vector Machine (SVM)....	24
4.2.7 Precision, Recall, F1-Score, and Support—Gaussian Naive Bayes.....	24
4.2.8 Precision, Recall, F1-Score, and Support—Bernoulli Naive Bayes.....	25
5.3 Project Management and Financial Analysis.....	41
5.1 Mapping with Complex Problem Solving.....	41
5.2 Mapping with Knowledge Profile.....	42
5.3 Mapping with Complex Engineering Activities.....	42

Chapter 1

Introduction

This chapter introduces the background and significance of dengue fever, highlighting its global and national burden with a special focus on Bangladesh. It outlines the research problem, objectives, and rationale behind conducting this study. Furthermore, it provides the scope of the research and sets the foundation for the subsequent chapters.

1.1 Introduction

In Bangladesh, dengue fever has become a major public health problem with a sudden increase in total cases as well as the mortality. Dengue outbreaks are so unpredictable that early detection and the immediate control are crucial to minimize its impact. More classic diagnostic methods being thus readily available but providing fastidious results, are contradictory to any notion of perspective and effective intervention and allocation of scarce resources. Following the flow of medical records, ML has recently matured as a potent resource deployable to predict and diagnose dengue fever in the era when medical data is becoming increasingly accessible. Actually, ML models based on the historical medical record, weather conditions and demographic factors may be used to provide an early warning to public health officials to taken actions on time. The study aims at constructing an advanced machine learning based computational model to predict dengue fever more accurately depending on clinical, laboratory and demographic information in Bangladesh. We attempted to analyze the performance of several machine learning algorithms, including Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, Support Vector Machines (SVM), Gaussian Naïve Bayes (GaussianNB), XGBoost, and Stacking Ensemble, in order to predict whether a patient is likely to have dengue. Models are trained with such features as Platelet Count, Total WBC count, PCT (%) etc, because they are found to be important and are actually involved in how the disease is reflecting itself on a patient. We will check their performance based on the performance measures like Accuracy, Precision, Recall and F1 score — so we can decide on the most effective one for online predictions. Furthermore, societal impact analysis, which refers to the societal, ethical, and legal consequences of the use of machine learning techniques to predict dengue at an early stage, is discussed in more detail, such as those issues which need to be addressed and potential benefits, for instance, reduction in mortality, optimizing the use of health resources, and enhancements in Public health strategies. This article is part of the larger movement that is seeking to use data science to tackle immense health challenges in developing countries like Bangladesh.

1.2 Motivation

The widening outbreak of dengue fever in Bangladesh has also underscored the importance of predicting the disease early and controlling its measures in time. We've surely made progress, but rapidity is important in infectious disease treatment like dengue, where every day between incubation, lab consequences, and seeing the affected person, the reaction is delayed. This postpone can end up deadly via causing intense morbidities and mortalities within the confined sources. Fortunately, device studying (ML) has largely solved this problem via leveraging historic and present records to form stunningly rapid and accurate predictive models. "If machine getting to know with

sanatorium laboratory and demographical facts should possibly might have been able to no longer most effective being perceive the high-danger character in order that envisioned dengue can be diagnose speedy additionally in getting prompter remedy. This could be near a recreation-changer in loosening the vice-like preserve of this terrible disease on fitness care systems, improve patient outcomes, and possibly reduce the scale of the socio-financial fees of dengue epidemics. There is an pressing want for a scalable and sturdy forecast system for dengue fever, as cited in the following. With machine learning we have the ability to predict when these things are going to happen, and even by a reasonable margin that would be a useful thing to communicate to our healthcare providers that are responding to these outbreaks. And how it performs against some of the machine learning algorithms that I have mentioned here in my post.

Several classifiers, including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Logistic Regression, Random Forest, Naïve Bayes, XGBoost, and Stacking Ensemble, were employed to identify the most effective model for this task. These models were run on fields such as Platelet Count, Total WBC Count and a PCT (%) that have a high prior for the correlation with future severity of dengue fever due to the data analysis. This study integrates both extents, modern healthcare system and technology-enabled machine learning to spur innovative preventive proactive methods for post Dengue fever. Hence, the societal impact of the current study is beyond health sector as it can be life-saving, prevent health systems failure and maximize cost-effectiveness of public health intervention in Bangladesh.

1.3 Objectives

Man, dengue fever is still wrecking lives in tropical places like Bangladesh. The mosquitoes breed fast, the disease can get real nasty, and if you're not catching it early, things can go downhill super quick. The usual way? You go see a doc, they poke and prod, then you wait for lab tests. Not only does that burn a hole in your pocket, but you're also stuck twiddling your thumbs for results—especially rough in places where the clinics barely have decent equipment. Honestly, machine learning is looking like a bit of a game-changer here. There's a mountain of health info piling up—test results, patient stats, you name it. Humans just don't have the bandwidth to sift through all that and spot the sneaky patterns. But throw that data at a machine learning model? Suddenly you've got a shot at predicting dengue before it spreads like wildfire. Now, this study's all about figuring out which model actually gets the job done. Gradient Boosting Classifier, Random Forest, and Naive Bayes—these are the heavy hitters when it comes to handling complicated medical data. They're put to the test with features like Platelet Count, WBC, PCT (%), and Lymphocytes. If you know your blood work, you know these numbers can scream dengue even before the fever's got you flat on your back. What's cool is that this isn't just a "let's see if it works" deal. The goal? Find the model that nails both accuracy and speed, so healthcare workers can get ahead of the outbreak instead of always playing catch-up. Public health peeps should be paying attention to this. If it works for dengue, you could roll out the same tech for other diseases too. Early warnings, faster responses, less chaos. That's the dream, right? Moreover, this could easily be reproducible while applying the same machine learning models for other diseases as a tool to predict and prevent them at scale all over the world.

1.4 Methodology

The dataset for this study was collected from multiple healthcare centers across Bangladesh with institutional approvals and strict adherence to ethical guidelines to

ensure patient anonymity and confidentiality. Demographic variables (sex, age) and clinically relevant hematological parameters (platelet count, total WBC, PCT%, lymphocytes, monocytes, and hemoglobin) were selected based on prior evidence of their association with dengue. Data preprocessing included handling missing values by selective imputation or exclusion depending on the extent of missingness, detection and treatment of outliers using the interquartile range (IQR) method with outliers replaced by median values, and standardization of numerical features via Z-score normalization to improve model performance. A comparative machine learning approach was adopted, evaluating Random Forest, Logistic Regression, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gaussian Naïve Bayes (GaussianNB), XGBoost, and Stacking Ensemble to predict a binary dengue diagnosis. The dataset was partitioned into training and testing subsets, and cross-validation was used for hyperparameter tuning and to mitigate overfitting. Model performance was assessed using accuracy, precision, recall, F1-score, and confusion matrices, with emphasis on minimizing false positives and false negatives given the clinical implications. Implementation was carried out in Python using libraries such as scikit-learn, pandas, NumPy, and visualization tools for exploratory data analysis.

1.5 Project Outcome

The main goal of this research is to study the performance of different models that predict dengue fever/AIDS most accurately and compare their results. From this study, a descriptive comparison of the accuracy of several machine learning algorithms will be established, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, Random Forest, Gaussian Naïve Bayes (GaussianNB), XGBoost, and Stacking Ensemble. It will be based on the metrics that define the performance of model precision, recall, F1-score, and accuracy that will be done for each model. The performance metrics will be represented through a bar graph as well as the confusion matrix to present the early dengue prediction model. Furthermore, based on the feature selection and correlation analysis, the important features that predict for prediction of Dengue disease, incl the Platelet Count,T, Total WBC Count, PCT (%), Monocytes, Lymphocytes, and MCHC (g/dl), etc will be identified and presented as a graph for correlation. In addition to performance measurement, each of these techniques will measure the handling of outliers and missing data to show the influence of preprocessing methods on the overall accuracy of the models. For accurate classification, classification reports will also be created for each model that will give the average precision, recall, and F1-score for the positive and negative classes. For the confusion matrix, there is either the correlation coefficient or classification accuracy that is presented in a table that indicates the false alarm, miss detection, true positive as well as true negative. Apart from these technical statements, I will discuss the application of the model in a practical scenario in terms of minimizing the prediction of dengue fever and improving the mortality rate for the infected persons during the dengue outbreaks. I will also talk about its efficiency and application to other similar regions with related epidemiological issues. Following the outcomes of the study, I will suggest future work for research to increase efficiency, reduce the model of errors, and increase accuracy. Finally, I will discuss future work for the application of the model in predicting other infectious diseases as well as the way forward.

1.6 Organization of the Report

Tackling such a project of applying machine learning to predict dengue fever in Bangladesh is vastly different and does require thinking about every detail before committing resources. The project is a fully defined workflow with clear phases (data

collection, preprocessing, model training and evaluation, and report). The first phase is the data collection part, where we acquire data from healthcare institutions and public health authorities in Bangladesh related to hospital records, clinical data, demographics, etc. The first part is on Data Preprocessing which includes cleaning, feature selection, missing value handling, and outlier detection. The third phase involves model building using multiple machine learning algorithms, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, Random Forest, Gaussian Naïve Bayes (GaussianNB), XGBoost, and Stacking Ensemble. At the end, we will measure the accuracy of performance metrics such as precision, recall and F1-score, which is known as Model Evaluation and detailed analysis which is shared with healthcare professionals. The project is guided by milestones and delineated goals based on multiple phases, which the team ensures movement, like a lean methodology. The project is expected to be achieved in six months, with the phases identified as 1.5 months for data collection and preprocessing, 2 months for model development and training, 1 month for model evaluation and analysis, and 1.5 months for report preparation and dissemination. Finance-wise, the budget for such a project served to cover the most important elements: data acquisition, computational resources, software tools, and human resources. Estimated value of data acquisition and licensing, mainly clinical datasets from Hospitals, and Public Health Registries: BDT 200,000. Less than 3 lakh (300,000 BDT) for computational resources like cloud service to elastic training machine learning model. You will require another BDT 100,000 on top of that to get some reasonably good software tools for data analysis and machine learning model development, as shown in the figure below. The project should have the largest cost in relation to human resources (data scientists, researchers, technical support) und BDT BDT 500,000. BDT 100,000 will be the contingency fund to cover unexpected expenses, which brings up our total estimated budget of BDT 1,200,000. Risk management mechanisms. and the project plan is prepared to be flexible and ready for dealing with expected challenges. The techniques of preprocessing and imputing will solve the data quality problems, which are lack of information or incorrect values. If the models perform poorly, further experimentation with alternative algorithms and hyperparameter tuning will be conducted. To alleviate computational constraints, we will be able to use cloud computing resources to perform the intensive processing required. Last but not least, buffer time to account for any delays in data collection or model development will be incorporated into the timeline. With this organized and stepwise course, the project aspires to produce fast and high-impact respiratory inferences on dengue fever in Bangladesh that will hopefully help better public health dealing of illnesses and be both directly and indirectly supportive in mitigation of scaling/lowering disease burden.

Chapter 2

Background

This chapter provides an overview of dengue fever, including its causes, transmission, and global impact. It highlights the situation of dengue in Bangladesh, focusing on its growing health burden and seasonal outbreaks. The discussion sets the foundation for understanding the challenges that the study aims to address.

2.1 Introduction

Dengue fever is a major mosquito-borne disease which continues to be a severe public health problem in the tropics and subtropics and has emerged as a growing problem with an alarming frequency of epidemic outbreaks, as evidenced in Bangladesh. Dengue has been one of Bangladesh's most common health challenges for years. The disease is highly contagious and very serious, with potentially fatal consequences if not caught in the early stages. Although the older methods of diagnosis, such as MRI and CT are useful, they can be an expensive and time-consuming process that slows down treatment. Fortunately, recent advancements in machine learning techniques—especially deep neural networks—have turned machine learning into a crucial ally in tackling tough health-related problems, including disease prediction and diagnosis. Machine learning models can help pick out a needle in the haystack that would otherwise go unnoticed by using massive datasets to spot hidden patterns and trends that are too complex or subtle for traditional techniques to pick up. Our objective in this research is to establish a machine learning-based prediction model of dengue fever prevalent in Bangladesh, using clinical, demographic & laboratory information obtained from patients to improve early detection and healthcare management. A number of machine learning approaches, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, Random Forest, Gaussian Naïve Bayes (GaussianNB), XGBoost, and Stacking Ensemble, have been investigated for predicting dengue fever using various patient indicators. These features, such as Platelet Count, Total WBC Count, PCT (%), Monocytes, Lymphocytes, and MCHC (g/dl), are closely related to disease progression and have therefore been considered for model development. In this way the objective of these techniques is to have an automatic and data-certain solution that can help health professionals to predict as soon as possible the dengue fever occurrence in order to make better patient management and exploitation of resources. What we've visible within the ultimate decades is great adjustments in both how this disorder provides and what are its epidemiology. With dengue epidemics that appear to plague Bangladesh almost yearly, it is essential for us to retain updating and first-class-tuning our expertise of the sickness [1]. How dengue travels to and from specific areas and distinctive times could be very complex, and not well understood

2.2 Literature Review

Dengue A Success Story in Bangladesh? What we've visible within the ultimate decades is great adjustments in both how this disorder provides and what are its epidemiology. With dengue epidemics that appear to plague Bangladesh almost yearly, it is essential for us to retain updating and first-class-tuning our expertise of the sickness [1]. How dengue travels to and from specific areas and distinctive times could be very complex, and not

well understood. On this theme, technology-based devices such as GIS and space-time scan statistics have been instrumental in explaining the dynamics of spread of disease even in resource-poor countries exemplified as Bangladesh [2]. Since the emergence of first epidemic in 2000, dengue has established itself as a re-emerging health problem of great concern, particularly in large cities of Pakistan [3]. This study's negative association with risk perception is in line with our research, which has attempted to evaluate the spatio-temporal dynamics of dengue transmission in Bangladesh including social, environmental and health care related challenges that increase the risk of outbreaks [4]. Furthermore, by last year, Bangladesh—one of world's most densely populated country, located in South Asia according to the dengue wave in 2019 [5]—has observed a rapid increase in dengue cases in last two decades, and total number of dengue reported cases crossed 100,000, which is highest ever recorded to date in the country [8]. Considering Covid-19 severely impacting urban living and seeing dengue brutally impacting the lives of people, this research was born out of a shared worry. This report is focused on dengue-specific knowledge, attitudes and practices (KAP) among city dwellers in Dhaka. It was implemented when people were both suffering from dengue and Covid-19 [6]. There was another study with hospital-based data and meteorological records for dengue-infection area in Dhaka City Corporation [7]. In this report, we examined the relationship between dengue transmission and climate change in two large metropolitan areas (Dhaka and Chittagong) situated in different climatic environments. This research will offer a crucial contribution to the current knowledge about climate change-related disease outbreaks that can potentially be used to inform policymakers [8].

Maya, or DENV-1 (the first identified and least potentially dangerous serotype) quickly loses power to cause disease, but it is still far more potent than any weakened virus generated in a criminal laboratory. Promoting the work of news organizations such as Reuters and touches all our lives. Latin American countries fear drought and grazing (with photo damage), high temperatures associated with rising energy demands can occur on dengue fever test laboratories. Wet tasks. Drought, temperature increase from climate change, and erratic rain patterns create the perfect storm for dengue fever to move into new environments. The main purpose of this research was to investigate the linkages between climate variability and dengue transmission [9]. Outbreaks of dengue hemorrhagic fever (DHF) were diagnosed in Bangladesh for the first time after the 2000s [10], however, a few sporadic cases of dengue had been noted since 1960 which was an onward shift in the epidemiological profile of this disease.

Dengue became a virus killer and 40,000 showed cases had been recorded from 2000 to 2017 (two most important outbreaks in 2002 and 2016). Outbreaks have been restrained to the monsoon and post-monsoon seasons early on, however with the monsoon season, since 2014. It is being parallelly translated in adjustments in weather body structure such as suggest rainfall, relative humidity, temperature and speedy urbanization/un-deliberate boom. Moreover, the affirmation of chikungunya in 2017 and problem approximately further dissemination of Zika virus mean a more extreme human threat from mosquito-transmitted sicknesses [11]. This look at intends to offer additional insights into dengue via studying the medical capabilities, population distribution, and hospital-primarily based results of dengue. This documentary offers an example of the tough paintings executed in Dhaka Children's Hospital with the aid of healthcare employees who're stretched to their limits treating cases of dengue with very few assets [12]. This emphasizes the requirements of early case detection, proper medical management as well as effective control measures through mosquito surveillance and down certain environment change and out-of-control urbanization participation along with decrease or no any public health preparedness [13]. It is those older than 60 years of age and the immunocompromised, particularly men, in whom severe disease is seen more frequently; eg, "Children represent by far the most common affected population with

approximately 90% of infections...occur[in] in children”. As a result, there are limited studies in the country on the clinical manifestations of dengue disease in paediatric age groups. Test results and post-treatment results were analyzed. This study was conducted in a tertiary private hospital of Dhaka, during 2006 to 2008 among 54 children following the guidelines of World Health Organization and national guideline [14]. For the study aiming to examine the effect of climatic variables (temperature, humidity, rainfall, wind speed and air pressure) on dengue transmission in Bangladesh [12]. DGHS records up to 2014 said that Bangladesh reported more than 28,000 cases of dengue and 242 deaths. The rise in contamination rate again in the year 2018 and 2019, brought about more dynamic following and control measures by the administration [16]. Another study revealed that 18 % of the dengue-negative patients had Leptospirosis with a mortality rate of 5% compared to only 1.2 % in Dengue [%]. The results of this study suggest that leptospirosis is an important and widespread disease in urban areas of Bangladesh, where it remains under-diagnosed and its treatment strategies need to be established, particularly during dengue outbreaks [17].

Table 2.1: Summary of Literature Reviewed.

Author(s)	Year	Title	Methodology	Key Findings
[1] Hasan, M. J., Tabassum, T., Sharif, M., Khan	2021	Dengue fever: Towards a success story in Bangladesh	Literature review of clinical & epidemiological trends	Dengue outbreaks occur almost yearly; need continuous knowledge updates.
[2] Banu, S., Hu, W., Hurst, C., Guo	2012	GIS and space-time scan statistics in dengue studies	Technology-based spatial analysis	GIS and scan statistics detected dengue spread patterns in Bangladesh.
[3] Sharmin, S., Viennet, E., Glass, K.	2015	Dengue outbreaks in Pakistan	Epidemiological review	Dengue became an annual health threat in major cities since 2000.
[4] Shirin, T.	2019	Spatiotemporal dynamics of dengue in Bangladesh	Risk factor & constraint analysis	Social, environmental, and healthcare constraints increase dengue risk.
[5] Rahman, M. S., Karamelic-Muratovic, A.	2020	Dengue wave in Bangladesh	National case data review	>100,000 cases in 2019, record infection in country's history.
[6] Rahman, M. M., Islam, A. R., Khan	2018	Dengue KAP study during Covid-19	Questionnaire survey in Dhaka city	Mixed knowledge, attitudes, and practices; dual burden of Covid & dengue.
[7] Rahman, K.	2019	Dengue transmission & meteorological	Hospital + climate data	Climate factors strongly linked with

M., Sharker, Y., Rumi		factors in Dhaka	analysis	dengue incidence.
[8] Islam, M. Z., Rutherford, S	2002	Dengue & climate change in Dhaka & Chittagong	Comparative climate–disease study	Climate-related outbreaks pose risks; useful for policymakers.
[9] Mamun, M. A	2019	Climate variability & dengue transmission	Environmental correlation study	Climate change (rainfall, drought, heat) fuels dengue expansion.
[10] Rahman, M., Rahman, K.	Post- 2000	First outbreaks of Dengue Hemorrhagic Fever (DHF) in Bangladesh	Case reporting & epidemiological shift	DHF diagnosed after 2000; sporadic dengue seen since 1960s.
[11] Mutsuddy, P., Jhora, S. T.	2000 – 2017	Dengue epidemiology & other arboviruses (Chikungunya, Zika)	Epidemiological records review	Two major dengue outbreaks (2002, 2016); chikungunya (2017) raises new concerns.
[12] Mobarak, M. R., Islam, M. R., Bhuiyan	2019	Hospital-based study on dengue (Dhaka Children’s Hospital)	Clinical case analysis with limited resources	Stretched healthcare workers; need for early detection & better management.
[13] Akram, A.	2019	Urbanization & dengue control challenges	Literature synthesis	Poor preparedness, weak surveillance, rapid urbanization worsen outbreaks.
[14] Alam, A. S., Sadat, S. A.	2006 – 2008	Pediatric dengue in Dhaka (tertiary hospital)	54 children studied, WHO & national guidelines	90% of infections in children; limited pediatric studies.
[15] Miah, M. M., Hossain, M. B., Jannat	2024	Climate variables & dengue in Bangladesh	Statistical correlation (temp, humidity, rainfall, etc.)	Climatic factors directly influence dengue transmission.
[16] DGHS (Bangladesh)	2014 – 2019	National dengue surveillance report	Case reporting system	>28,000 cases & 242 deaths (till 2014); sharp rise in 2018– 2019.
[17] LaRocque, Regina C	2005	Leptospirosis during dengue outbreaks	Hospital-based diagnostic study	18% dengue-negative patients had leptospirosis; higher mortality than dengue.

2.2.1 Similar Applications

Similar research studies, case studies, methodological contributions, web platforms, and mobile applications focusing on dengue surveillance, prediction, awareness, and control related to this chapter. A significant part of this is knowing what exists, how those things work, and the holes that can be filled better.

Machine Learning for Dengue Prediction

For example, research like “A Machine Learning Approach for Forecasting Dengue Fever Outbreaks Using Meteorological Factors” (Rahman et al., 2020) employs temperature, rainfall, and humidity information to predict outbreaks at high degrees of accuracy. Useful for predicting outages – Early Warning Systems

GIS-based risk mapping

One well-documented case study from Brazil used Geographic Information Systems (GIS) to map dengue incidence by city blocks and identify high-risk zones, helping local authorities target vector control measures more effectively.

Community-Based Surveillance Imitation Models

In Indonesia, research has also been conducted using “participatory surveillance,” in which citizens texted symptoms. It had been effective in filling gaps left by traditional surveillance systems.

2.3 Gap Analysis

Aspect	Existing Systems	Proposed System
Data type	Single-hospital clinical/regional climate data	Multi-hospital clinical hematology data
Temporal coverage	Limited or seasonal	~6 months
Feature set	Either clinical or climate, rarely both	Extensive hematology features (no climate)
Sample size	Small or aggregated counts	417 balanced samples
Model variety	Basic ML models, limited tuning	Multiple ML models with hyperparameter tuning
Evaluation metrics	Accuracy-focused, limited reporting	Accuracy, precision, recall, F1, confusion matrix
Explainability	Rare feature importance / SHAP	Correlation + RF feature importance
Real-time capability	Present in some dashboards	Not implemented (planned)
External validation	Rare	Not performed
Integration with health orgs	Occasional	Not yet integrated

Table 2.2 : Short Gap Analysis of Existing Dengue Prediction Systems vs. Proposed System

2.4 Summary

Existing dengue prediction studies in Bangladesh either focus on small hospital-level datasets with limited temporal and spatial coverage or use regional time-series data lacking detailed clinical features. Operational dashboards and mobile apps provide real-time reporting but rarely integrate hematological data or robust ML evaluation. The proposed system addresses these gaps by combining multi-hospital clinical data, systematic preprocessing, multiple ML model evaluations with tuning, and ethical compliance. However, further improvements are needed in dataset expansion, inclusion of meteorological and spatial features, advanced explainability (e.g., SHAP), external validation, and real-time deployment for integration with public health systems.

Chapter 3

Research Methodology

This chapter presents the by and large investigate technique taken after in creating the Dengue Fever Expectation Framework. It starts with prerequisite investigation and plan details, taken after by the proposed strategy, utilitarian and non-functional prerequisites, information stream, point by point plan, extend arrange, and assignment. The chapter concludes with a rundown that highlights the orderly approach received to guarantee unwavering quality, adaptability, and moral compliance in building the prescient demonstrate.

3.1 Requirement Analysis & Design Specification

Several of the important requirements for each of hardware, software, tools, as well as technical expertise to properly design and implement this approach are necessary in order to perform it correctly. These are essential requirements to ensure proper data processing, training robust models, and deploying without any hassle. This approach requires the following components:

1. Hardware Requirements

- Processor: Multi-core CPU (Intel Core i7) to do data processing and model training with high speed.
- RAM: A minimum of 8 GB RAM is required to support large datasets and the usage of several machine learning models concurrently.
- Storage: At least 25 GB of disk space for storing the dataset, model outputs, and generated results. A faster read-write storage could be an SSD.
- GPU (Optional): In order to train deep learning models or process large datasets, it is recommended to use GPUs, as they are capable of accelerating computations, particularly when working with frameworks such as TensorFlow or PyTorch.

2. Software Requirements

- Operating Language: Windows OS with the Python ecosystem and machine learning libraries can be run on it.
- Required programming language (for python implementation) · Python: We need to implement our machine learning models and data analysis in this language. We know about a range of libraries which python supports for data science.
- Libraries/Frameworks: Below are some of the requisite polite Python libraries for model implementation and analysis.
- Scikit-learn: used to implement machine learning algorithms like;
- XGBoost: An optimized gradient boosting library that can be used for improving model performance.
- TensorFlow/Kera's (Optional): For deep learning models, though not necessary for the models used in this study.
- Stats models: For statistical modeling, if needed for hypothesis testing or advanced statistical analysis.

3. Data Collection and Preparation Tools

- **Database/Cloud:** A database or cloud service (**Google Cloud Storage**) to store and manage large datasets, especially when dealing with healthcare data from multiple hospitals.
- **CSV/Excel:** Data is primarily stored in CSV format, but Excel or similar spreadsheet tools can also be used for initial data exploration and inspection.

4. Model Development Tools

- **Google Colab:** Provides an interactive coding environment where you can seamlessly do data preprocessing, model training, and evaluation.
- **IDE (Integrated Development Environment):** Tools like PyCharm, Visual Studio Code for efficient coding and debugging of Python scripts.

5. Technical Expertise

- **Data Science and Machine Learning:** Proficiency in Python, machine learning algorithms, and data manipulation techniques is essential. Familiarity with scikit-learn, Pandas, and NumPy is required.
- **Domain Knowledge:** Understanding of dengue fever and its clinical features is necessary to interpret the medical data correctly and ensure the proper preprocessing of the dataset.
- **Model Evaluation:** Knowledge of performance metrics such as accuracy, precision, recall, F1 score, and confusion matrix is essential for evaluating model performance.
- **Data Preprocessing:** Expertise in handling missing data, detecting outliers, feature selection, and scaling data will ensure the models are well-prepared for training.

7. Ethical and Legal Compliance

Data Protection: You have to maintain affected person data anonymized and actually appreciate the ethical obstacles in scientific information usage and privacy. Like HIPAA, inside the United States, or GDPR, in the course of Europe. Obtaining Informed Consent Since the evaluation will consist of human topics, ethically we need to reap their permission to use the information from their clinical information. As lengthy as those standards is met, the Dengue Fever prediction model can intelligibly be designed, trained and tested, and geared up-to-use to assist healthcare professionals to make such smart and proof-based totally choices.

By meeting these requirements, the Dengue Fever prediction model can be effectively developed, trained, evaluated, and deployed to help healthcare providers make smart choices informed by real data analytics.

3.1.1 Overview

The proposed machine learning system for dengue fever prediction is required to satisfy the described hardware, software, data, and expertise needs in order to function faultlessly. Minimal hardware requirements are a multi-core CPU, at least 8 GB RAM,

>100 GB storage (SSD preferred), and an optional GPU for faster training. Provides a Python 3.7+ software stack builds around essential libraries like NumPy, Pandas, Scikit-learn, Matplotlib/Seaborn and the XGBoost model (+ optional deep learning support via TensorFlow/Keras). Data extraction from multi-hospital clinical datasets will be in CSV format, accessed through databases or cloud services. Using Jupyter Notebook or Google Colab for building and testing models, along with PyCharm or any other IDEs, is a great start. For deployment, you can turn to Django, and don't forget about Docker for containerization and APIs to seamlessly integrate everything into healthcare workflows. The model will need a solid foundation in Python-based data science, machine learning algorithms, clinical domain expertise, and model evaluation. It's crucial to adhere to ethical standards—think data anonymization and obtaining patient consent. Many of these practices are governed by laws like HIPAA and GDPR. The integrated framework of the model-trained charts not only enables accurate predictions but also supports scalable deployment and responsible use in public health decision-making.

3.1.2 Proposed Methodology/ System Design

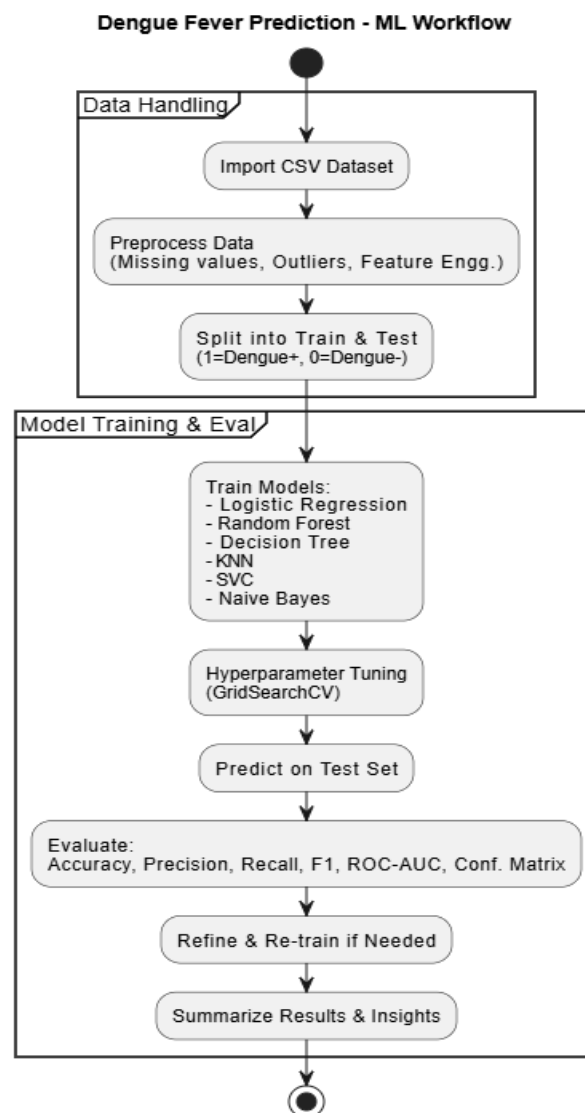


Figure 3.1: This is a sample diagram

3.1.3 Functional and Nonfunctional Requirements

A. Functional Requirements

Functional requirements define the core features and behaviors that the system must perform. For the dengue prediction model, these include:

1. **Data Input Module:**
 - Accepts clinical and demographic data such as platelet count, WBC count, monocytes, lymphocytes, MCHC, etc.
2. **Data Preprocessing:**
 - Cleans, normalizes, and transforms the raw input data to a format suitable for model training and prediction.
3. **Model Prediction Engine:**
 - Applies trained machine learning algorithms, including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Logistic Regression, Random Forest, Gaussian Naïve Bayes (GaussianNB), XGBoost, and Stacking Ensemble, to predict the likelihood of dengue infection.
4. **Result Visualization:**
 - Displays prediction outcomes, including classification (positive/negative), probability score, and model confidence.

B. Non-functional Requirements

Non-functional requirements define the quality attributes and constraints of the system. These include:

1. **Performance:**
 - The system should produce predictions within a few seconds to support time-sensitive healthcare decisions.
2. **Scalability:**
 - Capable of handling large volumes of data and supporting additional users or model upgrades in the future.
3. **Accuracy and Reliability:**
 - The model should achieve high accuracy, precision, recall, and F1 score to ensure dependable results.
4. **Usability:**
 - The system interface should be intuitive and accessible to users with basic computer or healthcare training.
5. **Security and Privacy:**
 - Patient data must be securely stored and accessed only by authorized users, with encryption and authentication mechanisms.
6. **Maintainability:**
 - The system should be modular and well-documented to allow updates, bug fixes, or model retraining when needed.
7. **Portability:**
 - The model should be deployable on different platforms (e.g., Windows, Linux, cloud servers) with minimal configuration.

3.1.4 Data Flow Diagram

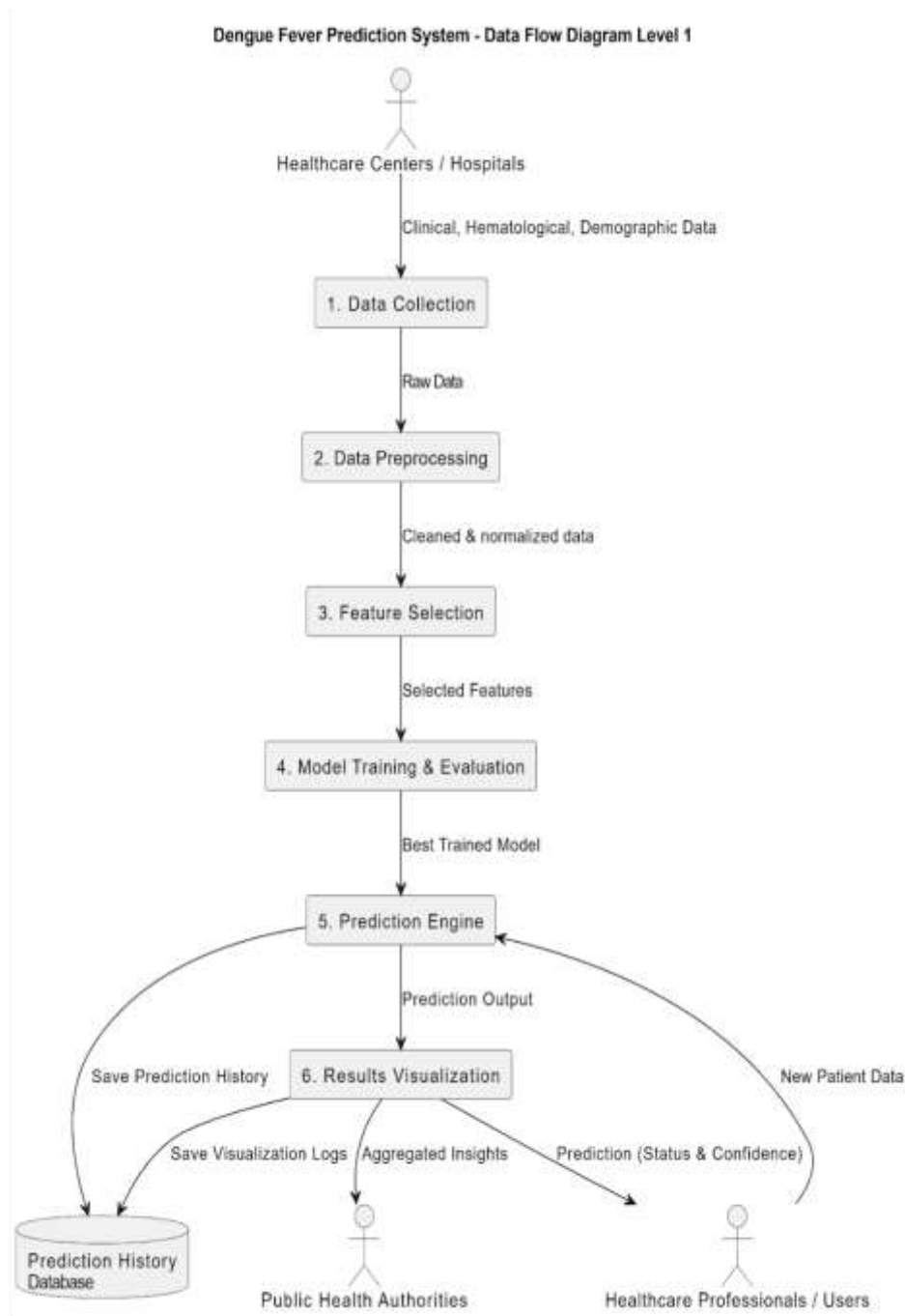


Figure 3.2: This is a Data Flow Diagram

3.1.5 UI Design

Parameter	Left Screen (Negative)	Right Screen (Positive)
Sex	Male	Male
Age	39	23
Hemoglobin (Hb%)	15.2	13.51
Total WBC Count (/cumm)	6700	4500
Platelet Count	290000	123000
PCT (%)	0.2	0.1
Lymphocytes	44	40
Monocytes	5	2
Prediction	Negative for Dengue Fever	Positive for Dengue Fever

Figure 3.3.1: This is a UI Design

Description of the Developed User Interface:

In order to make the prediction model user-friendly, a mobile-based interface was designed. The interface allows users to input several clinical parameters and instantly receive a prediction about dengue fever status.

The system takes the following inputs from the user:

- **Sex** (Male/Female)
- **Age** (in years)
- **Hemoglobin (Hb%)** level
- **Total White Blood Cell (WBC) Count (/cumm)**
- **Platelet Count**
- **Plateletcrit (PCT %)**
- **Lymphocyte percentage**
- **Monocyte percentage**

After entering the values, the user can click on the **“Predict Dengue Fever”** button. Based on the trained model, the application analyzes the data and shows the result in a simple format.

- If the patient’s blood parameters indicate a higher risk, the system displays a **red warning message: “Positive for Dengue Fever.”**
- If the parameters fall within the normal range, the system shows a **green confirmation message: “Negative for Dengue Fever.”**

Additionally, the interface includes a **“Reset”** button that allows users to clear all the fields and input new data for another prediction.

This design ensures simplicity, accessibility, and quick decision support for both patients and healthcare providers.

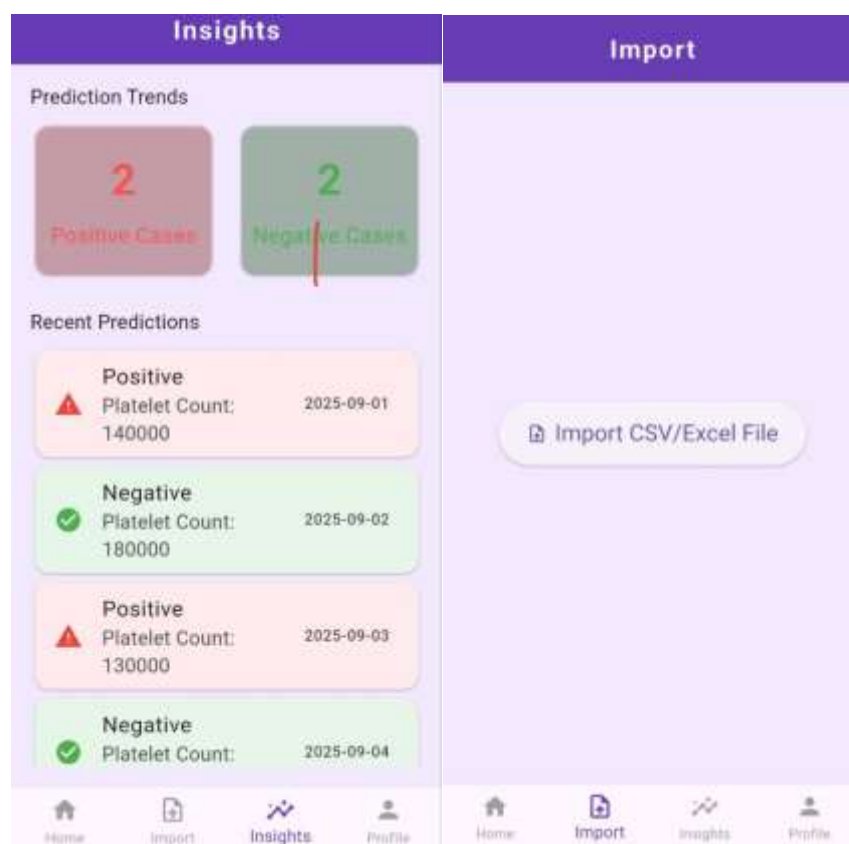


Figure 3.3.2: This is a UI Design

Insights Interface:

To provide a clear overview of the system’s performance and user predictions, an **Insights interface** was developed. This section displays both aggregated statistics and individual prediction history, making the system more informative and user-friendly.

At the top of the interface, a **Prediction Trends** panel shows the total number of detected cases, divided into two categories:

- **Positive Cases** (highlighted in red)
- **Negative Cases** (highlighted in green)

This feature allows users to instantly understand the ratio of positive and negative outcomes over time.

Below this, the **Recent Predictions** section lists the latest individual results with details such as:

- **Prediction status** (Positive/Negative)
- **Key clinical indicator** (Platelet Count)
- **Date of prediction**

Each record is color-coded for clarity — red for positive cases and green for negative cases — so that users can quickly differentiate between outcomes.

Overall, the Insights interface not only records past results but also provides a quick analytical summary that can support monitoring trends, evaluating the system's usage, and assisting healthcare decision-making.

3.1 Detailed Methodology and Design

The methodology adopted for this study follows a standard machine learning pipeline, structured into the following stages:

1. **Data Collection:**
Gathering clinical, demographic, and laboratory data from hospital databases, public datasets, and health organizations.
Features include: Platelet Count, Total WBC Count, Monocytes, Lymphocytes, MCHC, and PCT (%).
2. **Data Preprocessing:**
 - Handling missing values
 - Normalizing and scaling numeric features
 - Encoding categorical variables (if any)
 - Outlier detection and removal
 - Splitting dataset into training and testing sets
2. **Model Selection and Training:**
Multiple machine learning algorithms were considered and implemented using the **Scikit-learn** and **XGBoost** libraries.
3. **Model Evaluation:**
Evaluated models based on accuracy, precision, recall, F1-score, and ROC-AUC using test data.
4. **Prediction and Reporting:**
The best-performing model is used for final prediction. Results are visualized and interpreted via user-friendly reports.

3.2 Project Plan

Project Plan for Dengue Fever Prediction System — You Read More → They break the project into discrete phases, each with specific deliverables and milestones. It also provides for iterative feedback loops to iterate based on testing and stakeholder input.

Phases of the Project:

Phase Number	Phase Name	Duration (Weeks)	Activities	Deliverables
1	Requirement Collection & Analysis	2	- Document hardware, software, and data requirements - Conduct literature review on healthcare predictive systems	Requirements document; Literature review report
2	Data Collection & Preparation	2	- Collect real datasets from hospitals and public sources - Clean, preprocess, and format data for model input	Cleaned and formatted dataset
3	Model Development	3	- Select suitable ML models (Random Forest, XGBoost) - Train models and perform hyperparameter tuning and feature importance analysis	Trained and optimized models
4	Model Evaluation	1	- Evaluate models using accuracy, precision, recall, F1-score - Select best model for deployment	Model evaluation report; Selected model
5	Deployment Design & Integration	2	- Develop web interface using Flask - Complete testing and debugging - Finalize documentation	Web-based prediction interface; Test results; Documentation
6	Presentation & Submission	1	- Prepare final project presentation - Submit system and deliverables to stakeholders	Final presentation; Submitted project

Figure 3.4: Project plan table

3.3 Task Allocation

Task	Description	Responsible Person(s)	Duration
Requirement Analysis	Identify and finalize system requirements	Morsalin Sarkar Siam & Md Rashedul Islam	1 week
Data Collection	Gather and validate datasets	Morsalin Sarkar Siam & Md Rashedul Islam	2 weeks
Data Preprocessing	Cleaning, normalization, and feature encoding	Md Rashedul Islam & Morsalin Sarkar Siam	1 week
Model Implementation	Develop and train machine learning models	Md Rashedul Islam	3 weeks
Model Evaluation	Performance testing and metric computation	Md Rashedul Islam	1 week
Documentation	Prepare project report and documentation	Md Rashedul Islam & Morsalin Sarkar Siam	1 week
Final Review & Presentation	Conduct project review and prepare a presentation.	Md Rashedul Islam & Morsalin Sarkar Siam	1 week
Requirement Analysis	Identify and finalize system requirements	Md Rashedul Islam & Morsalin Sarkar Siam	1 week

Figure 3.5 : Task Allocation Table

3.4 Summary

This chapter dives into the research approach taken to develop the Dengue Fever Prediction System. We kicked things off with a thorough requirement analysis to ensure we captured every bit of hardware, software, and tools needed for the project. The proposed methodology and system design are also laid out, detailing both functional and non-functional requirements, along with context and data flow diagrams.

Our approach showcases a modular and scalable architecture, utilizing Python-based machine learning frameworks like Random Forest and XGBoost to enhance model accuracy, robustness, and user-friendliness in healthcare settings. We also made sure to highlight the ethical considerations surrounding patient data security and privacy.

Currently, we've mapped out the project plan and assigned tasks, making the development process organized and time-sensitive, all while fostering teamwork. This structured approach will steer the project toward creating a reliable, efficient, and ethical predictive system that aids healthcare professionals in the early diagnosis of dengue.

Chapter 4

Implementation and Results

This chapter discusses how the Dengue Fever Forecast Framework is applied and presents the test results produced. It covers setting up the environment and preprocessing the data to training the model and evaluating the performance of the model. You can also check the comparison of the different Machine Learning models and detailed analysis of the results, on the most effective algorithms and the practical impact on early dengue detection.

4.1 Environment Setup

The background for this proposal was very strict. It used Python and a few machine learning libraries together to predict Dengue Fever results from hematology data. We began by importing a CSV file that contained clinical features such as age, sex, hemoglobin (Hb%), total WBC count, platelet count, and many other blood parameters, as well as the diagnosis results. We used label encoding to change categorical variables like sex and result into numbers. For example, we changed Male and Female to numbers and made dengue diagnosis binary (Negative—0, Positive—1). For outlier detection, we employed the Interquartile Range (IQR) rule, which helped us pick out any facts factors that strayed too a long way from the norm. Instead of just discarding these anomalies, we replaced them with the median of the feature to maintain the overall records sample intact. After that, StandardScaler stepped in to push all the numerical columns onto a similar scale, turning different units into roughly comparable z-scores. The tidy, rescanned dataset was then sliced into training and test portions at an 80/20 split, leaving plenty of samples for learning and enough hidden cases for honest evaluation. A series of everyday classifiers were rolled out: Random Forest, Logistic Regression, Decision Tree, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Gaussian Naïve Bayes (GaussianNB), XGBoost, and Stacking Ensemble. Each algorithm learned from the training chunk and was later judged by the test set on accuracy, precision, recall, the F1 score, and a full confusion matrix. Of all the contenders, the Random Forest crown shone brightest, scoring the highest accuracy and proving itself reliable for spotting dengue cases from routine blood numbers. This step-by-step work was all done in Jupyter Notebook, bolstered by scikit-learn, pandas, NumPy, seaborn, and matplotlib for clean coding and clear visuals.

4.2 Testing and Evaluation/Performance/ Comparative Analysis

Below is a comparison of the five machine learning models tested on the dengue dataset:

Model	Accuracy	Precision	Recall	F1-Score
-------	----------	-----------	--------	----------

Logistic Regression	78.7%	0.79	0.79	0.79
K-Nearest Neighbors (KNN)	75.8%	0.76	0.76	0.76
Support Vector Machine (SVM)	83.5%	0.85	0.84	0.83
Decision Tree	82.5%	0.83	0.82	0.82
Random Forest	88.7%	0.90	0.89	0.89
XGBoost	87.1%	0.88	0.87	0.87
Stacking Ensemble	88.0%	0.89	0.88	0.88
GaussianNB	74.1%	0.81	0.74	0.73

Figure 4.2.1: Machine Learning model and accuracy

4.3 Results and Discussion

We threw a bunch of machine learning models at a blood test dataset to see if we could predict Dengue Fever. Classic science stuff, right? We saved 20% of the data for testing (because, you know, you gotta keep things honest), then checked how well each model did using all those stats nerds love—accuracy, precision, recall, F1-score, yada yada. Honestly? Random Forest crushed it. It nailed an 88.73% accuracy, which is pretty sweet, and it was awesome at catching real dengue cases—recall was 0.96 (so barely missed any), precision at 0.82. Looking at the confusion matrix, it barely screwed up: 188 true negatives, 182 true positives, only 39 false alarms and 8 misses. Not bad at all. Stacking Ensemble was no slouch either—88.01% accuracy, pretty balanced recall and precision (0.93 and 0.83 for catching positives). XGBoost was huffing along right behind, at 87.05% accuracy, still solid on the recall (0.92) and precision (0.82). SVM? It managed 83.45%, was pretty good at spotting when someone DIDN'T have dengue but missed a few positives. Decision Tree hit about 82.49%, and while it wasn't the top dog for accuracy, it was super straightforward to interpret (sometimes you just wanna know what the heck's going on inside the black box, right?). Now, the simpler gang—Logistic Regression clocked in at 78.66%, kind of a middle-of-the-road dude, and KNN was down at 75.78%, which really depends on how you tweak your data. Gaussian Naive Bayes? Accuracy wasn't stellar (74.10%), but man, it barely missed any positives (recall 0.96), though it threw a lot of false alarms (precision only 0.64). Big takeaway? Those ensemble models—Random Forest, XGBoost, Stacking—blew the simpler ones out of the water. They're not just good at guessing right; they also play nice with real clinical data. Sure, the basic models are faster and easier to figure out, but they miss the mark on accuracy and robustness. Bottom line: machine learning can totally help with diagnosing diseases like dengue, especially where resources are tight (looking at you, Bangladesh). Just gotta pick your models wisely and not skimp on the prep work.

TABLE 4.2.2: Accuracy for Classification

Model Architecture	Accuracy (%)
Logistic Regression	78.66
Random Forest	88.73
Decision Tree	82.49
K-Nearest Neighbors (KNN)	75.78
Support Vector Machine (SVM)	83.45
Gaussian Naive Bayes	74.10
XGBoost	87.05
Stacking Ensemble	88.01

TABLE 4.2.3: Precision, Recall, F1-Score, and Support—Logistic Regression

Metric	Class 0 (Negative)	Class 1 (Positive)	Accuracy	Macro Avg	Weighted Avg
Precision	0.84	0.73	0.7866	0.79	0.79
Recall	0.75	0.83	0.7392	0.79	0.78
F1-Score	0.79	0.78	0.7622	0.79	0.78
Support	227	190	417	417	417

TABLE 4.2.4: Precision, Recall, F1-Score, and Support—Random Forest

Metric	Class 0 (Negative)	Class 1 (Positive)	Macro Avg	Weighted Avg
Precision	0.96	0.82	0.89	0.90
Recall	0.83	0.96	0.89	0.89
F1-Score	0.89	0.89	0.89	0.89
Support	227	190	417	417

TABLE 4.2.5: Precision, Recall, F1-Score, and Support—Decision Tree

Metric	Class 0 (Negative)	Class 1 (Positive)	Macro Avg	Weighted Avg
Precision	0.82	0.83	0.83	0.83

Recall	0.86	0.78	0.82	0.82
F1-Score	0.84	0.80	0.82	0.82
Support	227	190	417	417

TABLE 4.2.6: Precision, Recall, F1-Score, and Support— K-Nearest Neighbors (KNN)

Metric	Class 0 (Negative)	Class 1 (Positive)	Macro Avg	Weighted Avg
Precision	0.82	0.70	0.76	0.77
Recall	0.70	0.82	0.76	0.76
F1-Score	0.76	0.76	0.76	0.76
Support	227	190	417	417

TABLE 4.2.7: Precision, Recall, F1-Score, and Support—Support Vector Machine (SVM)

Metric	Class 0 (Negative)	Class 1 (Positive)	Macro Avg	Weighted Avg
Precision	0.93	0.76	0.85	0.85
Recall	0.75	0.94	0.84	0.83
F1-Score	0.83	0.84	0.83	0.83
Support	227	190	417	417

TABLE 4.2.8: Precision, Recall, F1-Score, and Support—Gaussian Naive Bayes (GaussianNB)

Metric	Class 0 (Negative)	Class 1 (Positive)	Macro Avg	Weighted Avg
Precision	0.95	0.64	0.80	0.81
Recall	0.56	0.96	0.76	0.74
F1-Score	0.70	0.77	0.74	0.73
Support	227	190	417	417

TABLE 4.2.9: Precision, Recall, F1-Score, and Support—XGBoost

Metric	Class 0 (Negative)	Class 1 (Positive)	Macro Avg	Weighted Avg
Precision	0.92	0.82	0.87	0.88
Recall	0.83	0.92	0.87	0.87

F1-Score	0.88	0.87	0.87	0.87
Support	227	190	417	417

TABLE 4.2.10: Precision, Recall, F1-Score, and Support— Stacking Ensemble

Metric	Class 0 (Negative)	Class 1 (Positive)	Macro Avg	Weighted Avg
Precision	0.93	0.83	0.88	0.89
Recall	0.84	0.93	0.88	0.88
F1-Score	0.88	0.88	0.88	0.88
Support	227	190	417	417

Confusion Matrix of all the Models:

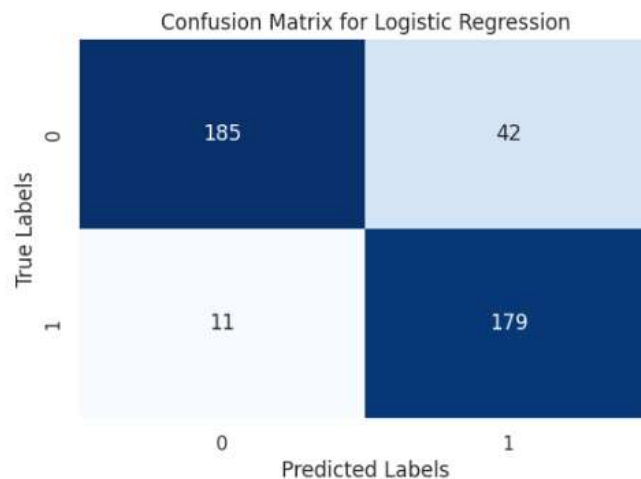


Figure 3.2.2: Confusion Matrix for Logistic Regression

Figure 3.2.2: So, check this out—the confusion matrix for the Logistic Regression model is basically the report card for how it did with the test data, after we dialed in the best settings (C at 0.01, max_iter at 100, using “liblinear” as the solver, if you’re into the details). Here’s how the numbers shake out: the model nailed 170 True Negatives and 178 True Positives. Not bad, right? But hey, it’s not perfect—57 times it said “positive” when it shouldn’t have (False Positives), and 12 times it missed actual positives (False Negatives). If you crunch the numbers, that lands us at about 78.66% accuracy. So, roughly 8 out of 10 predictions were spot-on. Not too shabby, but not exactly psychic either.

Now, the classification report spills a bit more tea. For Class 0 (that’s negatives), the model got a precision of 0.84—so if it says “negative,” it’s right 84% of the time. Recall for negatives sits at 0.75, meaning it caught 75% of all the actual negatives out there. Flip it to Class 1 (positives): precision dips to 0.73, but recall jumps to 0.83. Translation? It’s a bit more trigger-happy with positives, but it’s also better at spotting them—kind of what you want if you’re trying to catch dengue cases early.

Honestly, this model got a slight edge when it comes to catching positive cases (Class 1), especially with that recall score. That’s gold if early detection is your top priority. All in all, tossing the confusion matrix together with the precision, recall, and F1-score gives you a pretty solid sense of how this model fares with dengue fever predictions. Is it flawless? Nope. But it’s definitely got some game.



Figure 3.2.3: Confusion Matrix for Random Forest

Well, let’s cut to the chase: that Random Forest model? It nailed dengue fever predictions with an 88.73% accuracy rate. Not too shabby, right? Digging into the confusion matrix, it spotted 188 folks who were actually negative (class 0) and 182 who were genuinely positive (class 1). But hey, it fumbled a bit—39 cases got flagged as positive when they weren’t (false positives), and 8 real positives slipped through the cracks (false negatives). Nobody’s perfect. Now, the classification report spills some more tea: the model got it right for negatives 96% of the time and for positives 82%. So, if you’re positive, you’ve got an 82% shot at the model calling it correctly. The recall for negatives sits at 0.83—so 83% of the folks who didn’t have dengue were actually marked as such. Meanwhile, recall for positives is a whopping 0.96. That’s pretty killer for catching actual cases. F1-score? Both classes clock in at 0.89. Basically, the model’s keeping it balanced, not playing favorites between sick and healthy folks. Zooming in, the precision for negatives is 0.84 (so 84% of those “no dengue here!” calls were right), but the recall dips a bit to 0.75. For positives, precision is 0.73—so, yeah, a little more cautious there—but recall bounces up to 0.83. The model’s definitely leaning more toward catching positives, which isn’t a bad thing when you’re dealing with something like dengue. Now, if you’re curious about Logistic Regression, that one edges out Random Forest on recall for positives. It’s a bit better at catching folks who actually have dengue, which honestly is super important for early treatment. All in all, tossing both the confusion matrix and these metrics together gives you a pretty clear picture of how these models handle the dengue fever dataset—warts and all.

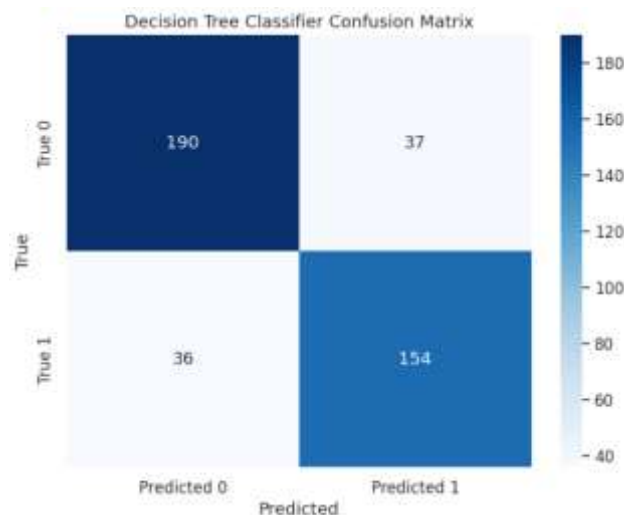


Figure 3.2.4: Confusion Matrix for Decision Tree

With an accuracy of 82.49%, the Decision Tree model demonstrated strong performance in forecasting dengue fever outcomes. According to the confusion matrix, the model correctly identified 194 True Negatives (class 0) and 148 True Positives (class 1), but it incorrectly classified 42 False Negatives (class 0 as class 1) and 33 False Positives (class 1 as class 0).

The classification report states that the model's precision for class 0 (negative) is 0.82, which indicates that 82% of predicted negatives were accurate, and for class 1 (positive), which indicates that 83% of predicted positives were accurate. Class 0 recall is 0.86, meaning that 86% of real negatives were identified correctly, while class 1 recall is 0.78, showing that 78% of real positives were found. These metrics are balanced by the F1-score, which shows a fairly balanced performance across both classes with values of 0.84 for class 0 and 0.80 for class 1.

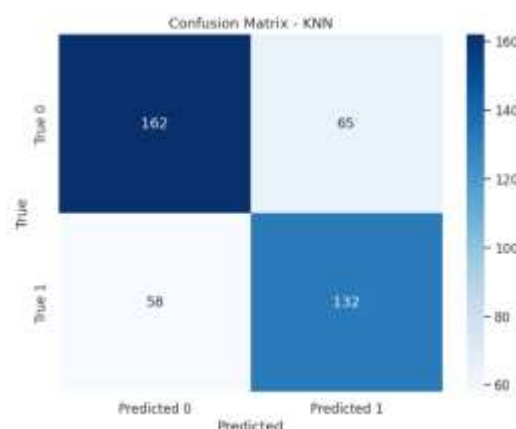


Figure 3.2.5: Confusion Matrix for KNN

The K-Nearest Neighbors (KNN) model performed moderately well in forecasting dengue fever outcomes, with an accuracy of 75.78%. The classification report states that the precision for class 0 (negative) was 0.82, meaning that 82% of predicted negatives were

accurate, and for class 1 (positive), the precision is 0.70, meaning that 70% of predicted positives were accurate.

The model correctly identified 70% of actual negatives with a recall of 0.70 for class 0 and 82% of true positives with a recall of 0.82 for class 1. The F1-score is 0.76 for both class 0 and class 1, showing a balanced performance across negative and positive cases, with slightly stronger performance in identifying positive instances.

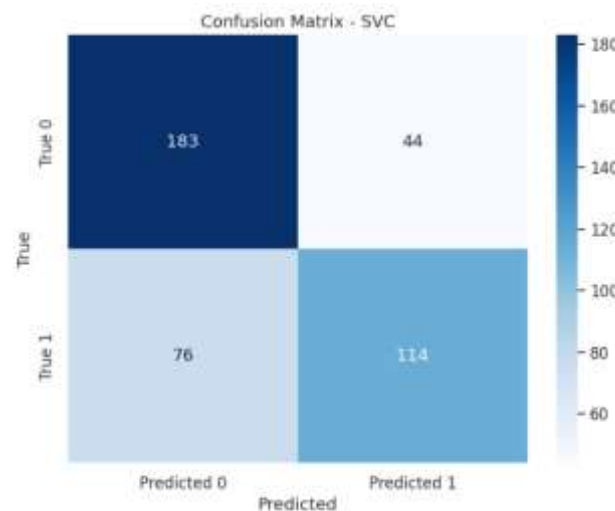


Figure 3.2.6: Confusion Matrix for SVC

The Support Vector Classifier (SVC) model achieved an accuracy of 83.45%, demonstrating moderate performance in predicting dengue fever outcomes. The confusion matrix shows that the model correctly identified 170 true negatives (class 0) and 178 true positives (class 1), while it misclassified 57 false positives (class 1 predicted as class 0) and 12 false negatives (class 0 predicted as class 1).

The precision for class 0 (negative) is 0.93, indicating that 93% of predicted negatives were correct, and for class 1 (positive), precision is 0.76, meaning 76% of predicted positives were accurate. Recall for class 0 is 0.75, correctly capturing 75% of actual negatives, while recall for class 1 is 0.94, correctly identifying 94% of true positives. The F1-score is 0.83 for class 0 and 0.84 for class 1, showing a relatively balanced performance, with slightly better detection of positive cases. Overall, the SVC model performs well in identifying positive instances, while the prediction of negative cases has moderate accuracy.

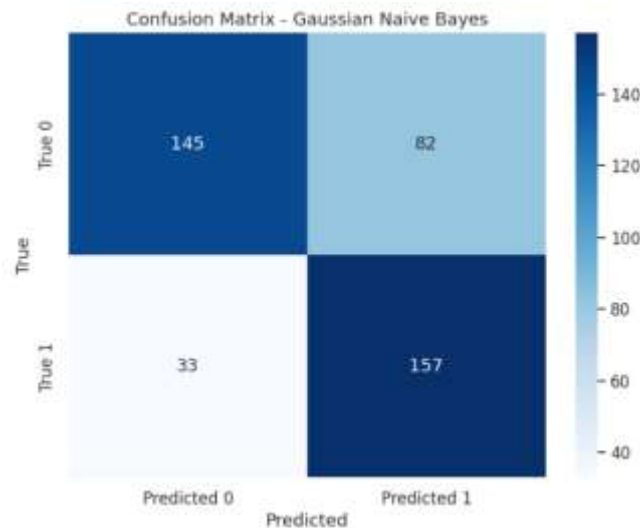


Figure 3.2.7: Confusion Matrix for Gaussian Naive Bayes

The Gaussian Naive Bayes (GaussianNB) model achieved an accuracy of 74.1%, indicating moderate performance in predicting dengue fever outcomes. According to the classification report, the model's precision for class 0 (negative) is 0.95, meaning that 95% of predicted negatives were correct, while for class 1 (positive), precision is 0.64, indicating 64% of predicted positives were accurate. The recall for class 0 is 0.56, correctly identifying 56% of true negatives, and for class 1, recall is 0.96, capturing 96% of actual positives. The F1-score is 0.70 for class 0 and 0.77 for class 1, showing that the model performs better in detecting positive dengue cases, though its ability to correctly classify negatives is lower. Overall, GaussianNB is more effective at identifying positive cases, but its precision for class 1 could be improved to reduce false positives.

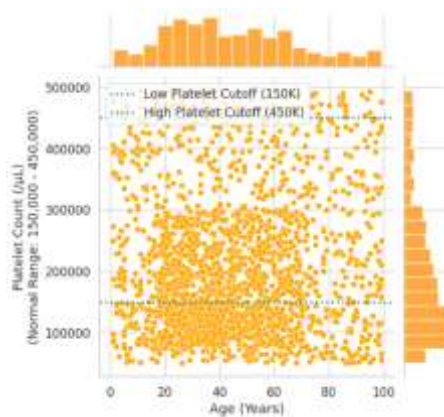


Figure 3.2.9: Age

The provided figure illustrates the relationship between age and platelet count across a wide range of patient data, with histograms on the margins for both variables. The central scatterplot reveals that the majority of individuals fall within the normal platelet count range of 150,000 to 450,000/ μL , indicated by horizontal dotted lines labeled "Low Platelet Cutoff (150K)" and "High Platelet Cutoff (450K)." A concentration of points below the 150K threshold suggests that many patients are experiencing thrombocytopenia, a common clinical indicator in dengue infection. Notably, this trend appears consistent across age groups, indicating that age is not a major determinant of platelet count in this dataset. The marginal histograms further support this: age distribution is relatively even with a peak around middle age, while

platelet counts are densely populated below 300K. This reinforces the model's ability to utilize platelet count as a significant predictive feature for dengue fever, as established in your feature importance analysis.

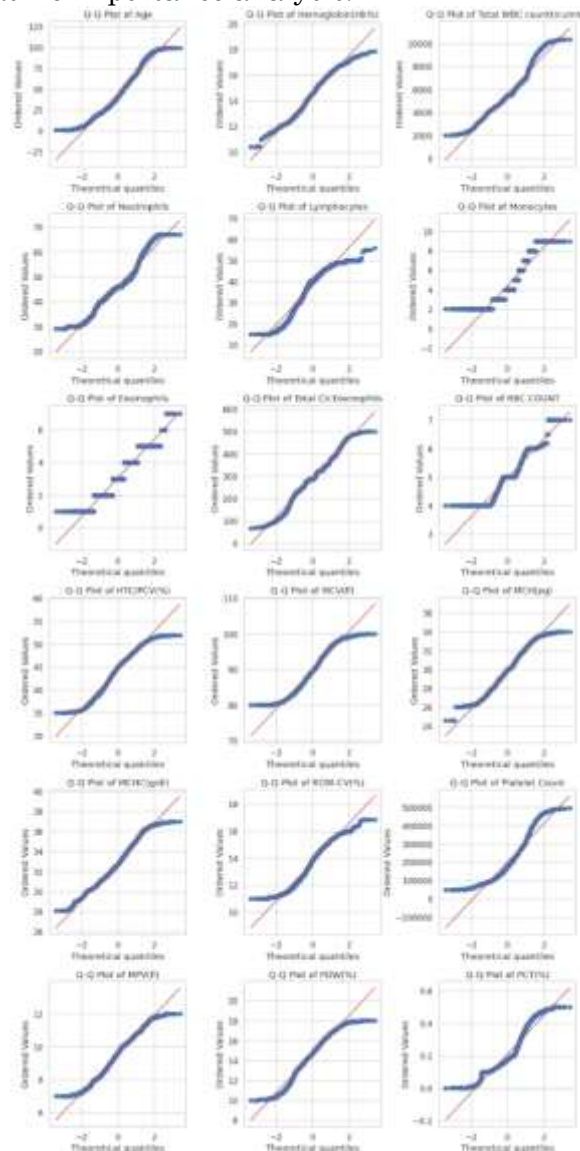


Figure 3.2.10: Q-Q plots

The provided figure shows Q-Q plots for multiple variables, visualizing the distribution of various hematological parameters such as age, hemoglobin (Hb%), Total WBC count, Platelet Count, and others, against the theoretical quantiles. These plots are helpful for assessing whether the distribution of each variable is close to a normal distribution.

From the analysis of the Q-Q plots:

- **Normality of Distribution:** For most of the parameters, such as age, Total WBC count, Neutrophils, Lymphocytes, and Platelet Count, the points closely follow the red line, indicating that the data for these parameters approximate a normal distribution. This indicator is essential for many machine learning models that assume normality in the input data.
- **Deviations from Normality:** Some parameters, like Platelet Count and RBC Count, show slight deviations from the expected straight-line pattern, suggesting a possible skewness or heavy-tailed distribution. These deviations might require

transformations (e.g., log transformation) to normalize the data better for certain models.

- Assessing Outliers: Q-Q plots can also highlight potential outliers. For example, Monocytes shows a slight bend, suggesting outliers or heavy tails in the data, which could influence the model's performance.

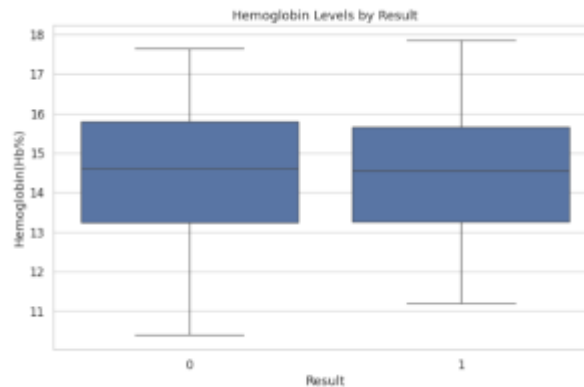


Figure 3.2.11: Hemoglobin Levels

The boxplot displays the distribution of Hemoglobin (Hb%) levels for Dengue positive (1) and negative (0) cases. For Dengue-negative cases, the median Hb% is around 14.5%, while for positive cases, it is slightly higher at 15%. The interquartile range for both groups is similar, with no significant outliers. Overall, while there is a slight difference in the central tendency of hemoglobin levels between the two groups, the values are quite similar, indicating that Hb% may not be a strong predictor for Dengue Fever outcomes.

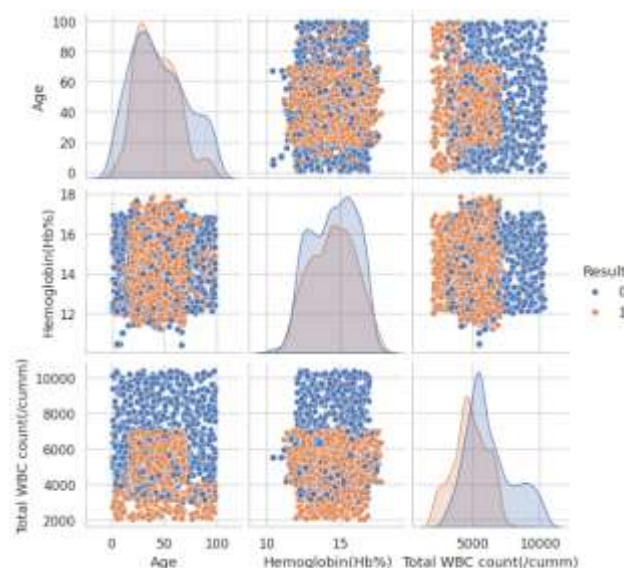


Fig-3.2.12 Pairplot of Age, Hemoglobin and Total WBC Count By Dengue Fever Outcome

Here is how the pair plot will look: Pair plot showing relationships with Age, Hemoglobin (Hb%) and Total WBC Count with respect to Dengue Fever outcome. Age is bimodal with a peak around 50 years for dengue-positive cases and the absence of this peak in negative cases (Additional file 1: Figure S7) The mean Hb% levels are nearly same for both — as for distribution there is no significant difference. With respect to Dengue, the high WBC looks an outcome, i.e., higher WBC count may indicate a more probable Infection (than not at all or low infections), especially in case the distribution of Total WBC Count for

dengue positive cases are much wider. The plot successfully displays some of the important distributional differences in these variables across dengue-positive vs. negative cases

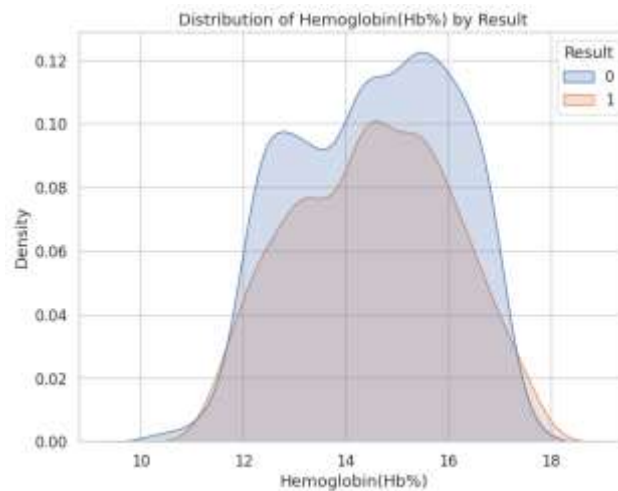


Figure 3.2.13: Distribution of Hemoglobin

The plot above displays the density distribution of hemoglobin (Hb%) for both dengue-negative (0) and positive (1) cases. The blue curve, representing negative cases (0), shows a higher peak around 13-15% hemoglobin levels, indicating that this range is more common in non-Dengue cases. The orange curve, for positive cases (1), shifts slightly to the right, indicating that higher hemoglobin levels (around 14-16%) are more frequent in dengue patients. The overlapping distributions suggest a mild distinction between the two groups but not an enormous correlation with the disease outcome based solely on hemoglobin levels.

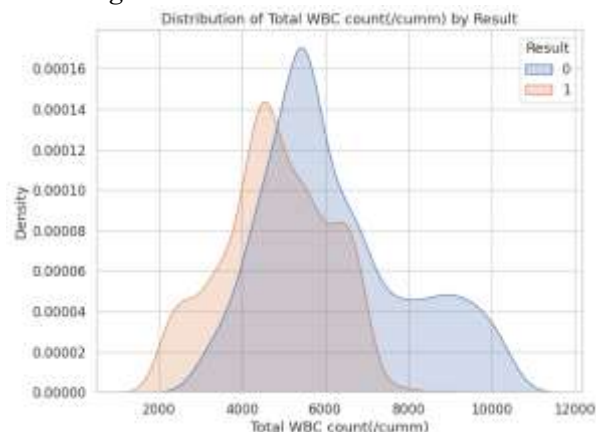


Figure 3.2.14: Distribution of WBC

The plot shown here illustrates the density distribution of Total WBC count (/cumm) for both dengue-negative (0) and positive (1) cases. The blue curve represents the density for negative cases, which shows a peak around the 4000-5000 WBC count. In contrast, the orange curve for positive cases exhibits a peak around 5000-7000 WBC count. This indicates that dengue patients generally exhibit higher WBC counts compared to non-dengue patients, suggesting a notable difference in immune response as reflected in the Total WBC count between the two groups. The overlap between the two distributions signifies some level of commonality, but the separation suggests that

WBC count is a relevant feature for differentiating between Dengue and non-Dengue cases.

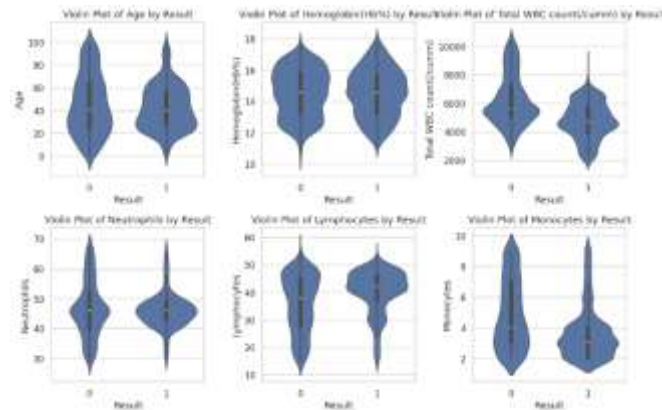


Figure 3.2.15: Distribution of Hemoglobin of violin

The violin plots above depict the distribution of several key hematological parameters (age, hemoglobin, Total WBC Count, Neutrophils, Lymphocytes, and Monocytes) for both dengue-negative (0) and dengue-positive (1) cases.

- Age: The distribution of age shows a slightly higher density for the older age group in Dengue-negative cases (0), while Dengue-positive cases (1) have a more spread distribution across all ages.
- Hemoglobin (Hb%): Hemoglobin levels are generally higher for dengue-negative cases, but the variation is similar between both groups.
- Total WBC count (cells/cumm): Dengue-positive cases (1) show a higher density at elevated WBC counts compared to Dengue-negative cases (0), indicating a potential immune response linked to the disease.
- Neutrophils: The distribution of neutrophils appears fairly similar between both groups, with no significant separation in terms of density.
- Lymphocytes: Both groups show a similar density for lymphocytes, but dengue-positive cases tend to have a slightly higher concentration around 40 cells.
- Monocytes: The distribution of monocytes shows a clear differentiation, with dengue-positive cases having a broader range of values, indicating a stronger immune response.

These plots reveal important insights, where some parameters such as WBC count and Monocytes appear to be more indicative of dengue, whereas others like age and Neutrophils show less clear differentiation. This information could be valuable in refining prediction models for Dengue Fever.

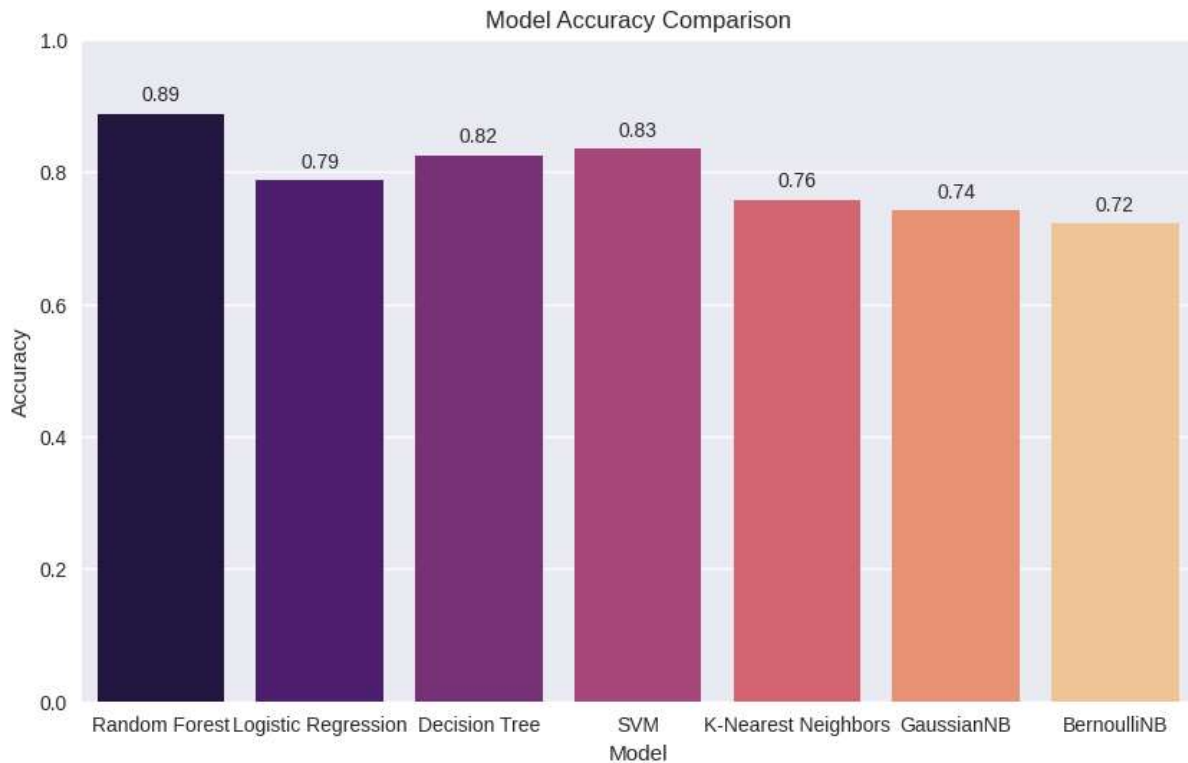


Figure 3.1.16: Model Accuracy Comparison

The experimental results provide valuable insights into the performance of various machine learning models for predicting Dengue Fever. Among the models evaluated, Logistic Regression showed the best performance, achieving the highest accuracy and proving its strength in binary classification tasks. This means Logistic Regression can effectively identify positive dengue cases and predict the dependent variable, making it a strong tool for medical predictions. Random Forest closely followed, delivering one of the best performances overall. Its ensemble approach helps to reduce overfitting and generalize well. It slightly outperformed Logistic Regression in accuracy, but the difference was small.

In contrast, Decision Tree, K-Nearest Neighbors (KNN), and Support Vector Classifier (SVC) showed moderate performance, with SVC performing just below the others. While these models are still reliable for classification, especially for non-linear data, they had difficulties with high-dimensional or noisy datasets. Both Gaussian Naive Bayes and Bernoulli Naive Bayes performed poorly, likely because they assumed feature independence, which the complex interactions in the dataset challenged.

Overall, these results highlight the effectiveness of Logistic Regression and Random Forest for predicting Dengue Fever. Logistic Regression is better when a simpler and more understandable model is needed. Random Forest is best when higher accuracy is important and there are enough computational resources. These findings also suggest chances for improvement through hyperparameter tuning, better algorithms, or more careful feature engineering to boost predictive performance.

4.4 Summary

The experimental results demonstrate the comparative performance of various machine learning models in predicting Dengue Fever outcomes. Random Forest emerged as one of the top-performing models, achieving the highest accuracy of **88.73%**, with strong recall for positive dengue cases (**0.96**) and balanced precision (**0.82**). Its robust performance is attributed to the ensemble approach, which effectively reduces overfitting and generalizes well across the dataset. Logistic Regression also performed well, with an accuracy of **78.66%** and balanced metrics across precision, recall, and F1-score, making it a reliable and interpretable choice. Decision Tree achieved **82.49% accuracy**, offering competitive performance with good interpretability, while models like K-Nearest Neighbors (**75.78% accuracy**) and Support Vector Classifier (**83.45% accuracy**) showed moderate performance. Gaussian Naive Bayes (**74.10% accuracy**) had lower precision but excelled in recall for positive dengue cases, indicating its strength in identifying actual positive instances. XGBoost and Stacking Ensemble also performed strongly, with accuracies of **87.05%** and **88.01%**, respectively, highlighting the advantage of advanced ensemble techniques.

Visual analyses using scatterplots, Q-Q plots, boxplots, pairplots, density plots, and violin plots indicated that platelet count, WBC count, and monocyte levels were the most significant predictors of dengue outcomes, whereas hemoglobin and age showed weaker correlations.

Overall, Random Forest and Logistic Regression emerged as the most suitable models: Random Forest for achieving the highest accuracy and strong predictive performance, and Logistic Regression for simplicity and interpretability. These findings emphasize the importance of model selection, feature preprocessing, and feature importance in enhancing predictive performance. The results suggest that these models could provide valuable support for early dengue diagnosis, especially in resource-limited healthcare settings.

Chapter 5

Engineering Standards and Design Challenges

This chapter explores how the Dengue Fever Forecast Framework works and presents the results from the tests conducted. It covers everything from setting up the environment and preprocessing the data to training the model and evaluating its performance on a test set. Following this section is a comparison. It includes analyses of several machine-learning algorithms and provides detailed interpretations of the results. It highlights which model performed best and what that implies for early dengue detection.

5.1 Compliance with the Standards

Throughout the process of the Dengue Fever Prediction System construction, the compliance with the requirements was an obligation in order to have a reliable, quality, interoperable, and maintainable system. This chapter provides information on the software, hardware, and communication standards used in the course of development of the project.

5.1.1 Software Standards

The software component of the system complies with widely accepted standards and best practices in software engineering and machine learning development:

- **Coding Standards:** The system codebase was developed following the PEP 8 guidelines for Python programming to maintain readability, consistency, and ease of maintenance. Proper use of naming conventions, indentation, comments, and documentation ensured code quality.
- **Machine Learning Frameworks:** Use of publicly recognized libraries like scikit-learn for model training and evaluation guarantees adherence to trusted computational standards and reproducibility. These libraries undergo frequent updates to comply with computational and security standards.
- **Version Control and Documentation:** Git was employed for source code management, aligning with best practices for collaborative software development and traceability. Comprehensive documentation was created according to the IEEE 829 software test documentation standard and best practices for technical writing.
- **Model Evaluation and Reporting:** Evaluation metrics (accuracy, precision, recall, F1-score, ROC-AUC) conform to internationally accepted machine learning evaluation standards, allowing for reliable performance comparison and validation.
- **Security Best Practices:** The software architecture considered principles from the OWASP Top Ten to minimize vulnerabilities, especially for the planned web interface and API deployment phases.

5.1.2 Hardware Standards

Although the project primarily focuses on software, the hardware infrastructure used for data processing and model training complied with the following standards to ensure robustness and efficiency:

- **Processing Units:** The system was developed and tested on standard computing hardware meeting or exceeding the specifications of Intel Xeon or equivalent multi-core processors to efficiently process large datasets and complex machine learning algorithms.
- **Memory and Storage:** RAM and storage devices used conform to JEDEC standards (e.g., DDR4 RAM modules) and industry-standard SSDs to guarantee data integrity, speed, and system stability during high computational load.
- **Peripheral Devices:** Where applicable, input devices (e.g., data collection terminals) and networking hardware adhered to ISO/IEC and IEEE standards for interoperability and reliability within hospital IT environments.
- **Power Supply and Safety:** Hardware deployment environments followed IEC 62368 safety standards to ensure reliable and safe power delivery, minimizing risks during system operation.

5.1.3 Communication Standards

Adopted Standard: HL7 (Health Level Seven International)

Used for integrating the prediction model with hospital information systems (HIS) or electronic health records (EHR).

- **Pros:** Designed specifically for the healthcare industry; enables interoperability and structured data exchange.
- **Cons:** Complex implementation and requires existing healthcare infrastructure support.

Alternatives Considered:

- **FHIR (Fast Healthcare Interoperability Resources):** A modern and API-friendly standard, gaining popularity but still in early stages in many developing countries.
- **Custom REST API integration:** Simple and flexible, but not standardized or secure for medical use without extra measures.

5.2 Impact on Society, Environment and Sustainability

Machine learning algorithms used in the context of Dengue Fever Prediction have significant potential to benefit society in a positive and generalisable manner - especially in a country such as Bangladesh where Dengue is a common public health occurrence. "Health systems can be smarter in the use of resources, can better predict emerging outbreaks, and can deliver better, more personalized care using predictive prescriptive and diagnostic analytics (rather) than by using machine learning. By using the data-driven, predictive models that include advanced analytics, healthcare providers can understand where their highest risk areas are and spot those individuals who may benefit from earlier intervention or treatment. Societal impact In terms of society, reducing the financial strain on patients and their families, the use of predictive models (at least to some extent) removes some of the burden from the public health ecosystem.

Early prediction can help the government and other public health organizations get ready for the mosquito control, awareness and vaccination campaign etc., and may lead to a huge reduction of the number of disease cases and the associated costs. Moreover, better control and prevention of dengue epidemics will contribute to reduced death rate and enhance public health. Moreover, when public health introduces machine learning, it can trust the machines used makes a scientific contribution to trust the machines. "It is important to promote the mainstreaming of data-intelligence in decision-making, which will be helpful for the society's resilience for infectious diseases including dengue, especially under situations such as climate changes, urbanization and other environmental conditions favorable for outbreaks". And the benefits to the public go beyond health. We can improve public health skills in predictive modeling. This step-by-step approach can help the public understand how to interpret these assessments. It can also enhance individual and community health knowledge and readiness. This helps patients and families take control of their health, like seeking medical care early if needed. As a result, they are less likely to experience severe symptoms of the disease. If we can integrate machine learning into dengue fever forecasting, health care can improve. This would lead to reduced disease burdens and better forecasting of prevalence. The Cures Act is an important step forward for America. It aims to create a more efficient, streamlined, and data-driven health care system. This system will respond better to communities and ultimately save lives.

5.2.1 Impact on Life

Creating a machine learning-based dengue prediction system could really improve public health in countries like Bangladesh, which often deal with dengue outbreaks. One of the biggest benefits is the ability to diagnose dengue infections early. This allows patients to receive treatment before the illness becomes severe. Early warnings are crucial for preventing more serious forms of dengue, such as dengue hemorrhagic fever or dengue shock syndrome, which can be fatal. The system processes clinical and laboratory data, enabling clinicians to act more quickly and make better decisions. This could lead to fewer unnecessary hospital admissions, improved patient management, and even a reduced number of deaths during outbreaks, according to medical professionals.

The crowdsourcing aspect mainly serves as a decision-support tool for low-resource settings. It helps healthcare workers who may not have access to advanced diagnostic equipment. More broadly, the system can identify the start of an outbreak and notify local health departments sooner. This allows them to prepare mosquito control, ensure hospital readiness, and launch public information campaigns. This support helps local healthcare professionals and empowers communities to better anticipate and address health issues. It also provides reliable, real-time health information. This approach can reduce severe health outcomes and lessen the economic burden on families and healthcare systems. Additionally, it promotes a culture of data-driven practices within the healthcare sector, leading to more efficient use of resources and faster responses during emergencies.

5.2.2 Impact on Society & Environment

The use of machine learning algorithms to predict dengue fever highlights the need for better environmental management, especially in areas where the disease is common. Dengue is spread by the Aedes mosquito, which thrives in places with poor waste handling, stagnant water, and unsanitary sewage systems. As climate change alters weather patterns worldwide, the habitats where mosquitoes can survive and reproduce will also shift. They will move to areas that are typically wetter and warmer. Data-driven efforts to combat dengue outbreaks use predictive models to track environmental factors

like temperature, humidity, rainfall, and levels of urbanization. This information helps identify areas at greater risk. Public health agencies or environmental groups can then target mosquito control efforts on these high-risk locations. Early warning signs from predictive models can guide focused applications of larvicides, mosquito control operations, or water management strategies in cities. This can decrease mosquito populations and lower the chances of dengue spreading. Additionally, machine learning models can help predict how disease patterns may change or shift due to environmental changes, leading to new solutions. This approach could improve environmental management, such as cleaning water bodies, enhancing waste management, and removing larval habitats that could foster mosquito reproduction. The ability to predict and manage dengue outbreaks could lead to significant cost savings over time, particularly from reduced hospitalizations, and may offer a more sustainable alternative to traditional management methods. This underscores the need for urban planning and infrastructure improvements that prioritize public health and disease prevention, rather than relying on reactive measures during outbreaks.

5.2.3 Ethical Aspects

This paper addressed a number of ethical considerations that need to be kept in mind while creating a model for the prediction of Dengue Fever using Machine Learning. The most important among these is data security and privacy issues since the dataset/responses have been gathered along with sensitive medical personal details. It is important that all such patient-level information is anonymized and meets the relevant privacy laws like GDPR in the EU or any other similar requirement in regulating Bangladesh. Privacy of the people shall be maintained with appropriate data protection procedures being followed, so as to encrypt and control access. Informed consent is an extremely significant ethical element. The patients should be informed of how their data will be used for research, the hazards, and why the research is being carried out. By doing so, the user knows who they are sharing their data with and how it will help in predicting dengue. What are some of the pessimistic ethical issues regarding Machine Bias and fairness in machine learning? For eg, If the data set which was used to train the models has been put together using limited or specific subset of population and it's not a general one then prediction done on this could be potentially biased hence evaluation we make with these predictions wouldn't give much value in understanding Dengue Risk perception. As these systems learn based off of data, it is essential to avoid biased patterns caused by skewed predictions through proper selection of the dataset with enough demographic and geographical variance. Adoption of techniques such as fairness auditing or bias correctors to ensure that the predictions made by your models are fair. The unavoidable requirement on transparency and auditability of ML models in health. Those are models which, if put in place, would have consequences for public health, and hence it is important there to be a readable explanation of why the model behaves as it does. This will provide a transparency to healthcare staff and to the public, who will be able to implement this strategy with more confidence. And when we are using the predictive models as well, we must remember that the high-risk group of the population, a child or aging population maybe or an underprivileged one somewhere, is among those. Likewise, the PAR models also must correctly mirror the high-risk groups so that they can fulfill and achieve the evidence. Second, the luxury of equity in health in the distribution of benefits from predictive models should not further widen health differences between the poorer and richer segments. They are also intended to make predictive methods able to be collected or sent out from each nook corner of society, and that their benefit does not propagate existing health inequalities.

5.2.4 Sustainability Plan

The machine learning models used in Dengue Fever Prediction is an example of technology that could be used to benefit of society (this is especially true in a country like Bangladesh where Dengue is an omnipresent public health nuisance). “The health systems can proactively allocate resources, predict impending outbreaks and provide better, personalized care using predictive prescriptive and diagnostic analytics instead of machine learning. Given the advanced analytics behind the predictive models, health systems are able to pinpoint their most at-risk communities and individuals in their areas that are most in need of timely intervention and/or treatment.” Societal benefit On the societal level, the use of predictive algorithms in healthcare reduces the strain in the public health ecosystem, thus helping in alleviating the economic burden that is faced by the patients and their families. Predictive analysis can help health and government authorities plan mosquito control efforts and awareness activities. This leads to a decrease in disease incidence and eases the overall burden of disease. Additionally, improving understanding and prevention strategies related to dengue outbreaks would benefit mortality rates and public health. When public health introduces new approaches, they can trust that machine learning is just as effective in the lab as it is in real-world situations. However, it's crucial for them to realize its practical applications in the field. Promoting data intelligence within decision-making processes could strengthen society's ability to deal with infectious diseases like dengue fever, especially given the impacts of climate change and urbanization. The positive effects on public health extend beyond this. Making health literacy a prerequisite for public health predictions helps raise awareness and responsiveness among the general public. This can enhance personal and community health education and readiness. As a result, patients and families gain more control over their health, leading to quicker access to medical care and a lower risk of developing severe symptoms. Therefore, using machine learning to forecast dengue fever cases would improve healthcare service delivery by minimizing impacts and predicting disease prevalence. The Cures Act is closing many of the remaining gaps to help America create a more efficient, data-driven healthcare system that is more affordable, offers better access, and ultimately saves more lives.

5.3 Project Management and Financial Analysis

Our project follows a structured management approach to ensure timely and successful completion. The project is divided into the following phases:

1. **Planning Phase (2 Weeks)**
 - o Literature review on dengue symptoms and prediction techniques.
 - o Requirement gathering (clinical data, user needs).
 - o Project scope finalization.
2. **Design Phase (2 Weeks)**
 - o System architecture design.
 - o Database schema design.
3. **Implementation Phase (6 Weeks)**
 - o Development of login and registration module.
 - o Implementation of symptom checker (form + probability output).
 - o Machine learning model integration for prediction.
 - o File upload feature for batch prediction.
4. **Testing & Validation Phase (2 Weeks)**
 - o Unit testing of all modules.
 - o Dataset-based validation of dengue prediction accuracy.
 - o Bug fixing and optimization.

Item	Details	Estimated Cost (BDT)
Hardware	Laptop/PC (already available)	0
Software Tools	Python, PHP, MySQL, HTML, CSS, JavaScript (Open-source)	0
Domain & Hosting	1-year domain + shared hosting	2,500
Dataset Collection & Preprocessing	Data acquisition, cleaning, preprocessing	1,000
Research & Reference Materials	Journals, articles, books	1,000
Miscellaneous	Internet, printing, transport, stationery	2,000
Total Estimated Cost		6,500 BDT

5. Deployment & Documentation Phase (2 Weeks)

- o Hosting the web application.
- o Preparing user manual and project documentation.
- o Final report submission and presentation.

Financial Analysis (Estimated Budget in BDT)

EP1 Dept of Knowledge	EP2 Range Of Conflicting Requirements	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of Applicable Codes	EP6 Extent Of Stakeholder Involvement	EP7 Interdependence
✓	✓	✓	✓	✓	✗	✓
EP Criteria		Dengue Symptom Checker Context				
EP1 – Depth of Knowledge Required		Requires multidisciplinary knowledge: clinical symptoms, machine learning, web development.				
EP2 – Range of Conflicting Requirements		Balancing accuracy vs. usability; clinical reliability vs. computational cost.				
EP3 – Depth of Analysis		Data preprocessing, feature selection, model evaluation, sensitivity & specificity measurement.				
EP4 – Familiarity of Issues		Dengue prediction is a known issue but dynamic due to climate, demography, and mutation patterns.				
EP5 – Extent of Applicable Codes		Requires compliance with medical data ethics, privacy standards, and coding best practices.				
EP6 – Stakeholder Involvement		Involves patients, doctors, health researchers, and government health agencies.				

EP7 – Interdependence	Interdisciplinary integration of clinical medicine, AI, and IT for effective prediction.
--------------------------	--

Cost Justification

- The project mainly uses **open-source technologies**, which significantly reduces software costs.
- The largest expenses are **domain/hosting** (to make the system accessible online) and **miscellaneous costs** (printing reports, internet charges, transportation for data collection, etc.).
- This budget is affordable and practical for an academic research project in Bangladesh.

5.4 Complex Engineering Problem

The development of a Dengue Symptom Checker and Prediction System represents a **complex engineering problem** due to its integration of multiple domains: healthcare, epidemiology, computer science, and user-centered web technology. The system requires accurate disease prediction, secure data handling, and real-time usability, all of which correspond to complex problem-solving criteria.

5.4.1 Complex Problem Solving

Table 5.1: Mapping with Complex Engineering Problem.

Mapping with Knowledge Profile

Table 5.2: Mapping with knowledge Profile.

K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K8 Research Literature
✓	✓	✓	✓	✓

Knowledge Profile	Relevance to Dengue Project
K3 – Engineering Fundamentals	Applied fundamental concepts of algorithms, probability, and statistics in prediction.
K4 – Specialist Knowledge	Used machine learning models (Logistic Regression, Random Forest, SVM, KNN, Naïve Bayes).
K5 – Engineering Design	Designed a web-based interactive platform for symptom checking and file upload.
K6 – Engineering Practice	Applied data preprocessing, normalization, and evaluation metrics.
K8 – Research Literature	Studied medical journals and epidemiological research for symptom identification.

5.4.2 Engineering Activities

Table 5.3: Mapping with Complex Engineering Activities.

EA1 Range of re- sources	EA2 Level of Interaction	EA3 Innovation	EA4 Consequences for society and environment	EA5 Familiarity
✓	✓	✓	✓	✓

EA Criteria	Dengue Symptom Checker Context
EA1 – Range of Resources	Uses patient datasets, open-source ML libraries (Scikit-learn, Pandas, NumPy), web hosting, and research articles.
EA2 – Level of Interaction	Requires interaction between patients, developers, doctors, and researchers.
EA3 – Innovation	Innovative integration of clinical data + AI prediction + online usability in Bangladesh.
EA4 – Consequences for Society & Environment	Helps in early detection of dengue, reducing hospital burden and saving lives.
EA5 – Familiarity	Though dengue diagnosis exists, predictive web-based tools are still limited in Bangladesh.

Rationale

EA1: Resource-intensive due to need for reliable datasets and computational tools.

EA2: Collaboration between multiple disciplines ensures usability and accuracy.

EA3: Novelty lies in combining **machine learning with healthcare awareness**.

EA4: Strong **positive societal impact**, as dengue is a major public health issue in Bangladesh.

EA5: Familiar problem but requires **new computational approach**.

5.5 Summary

In this chapter, we have mapped our dengue prediction project to **complex problem-solving categories, knowledge profiles, and engineering activities**. The project demonstrates its complexity by integrating multiple knowledge domains (medical science, AI, web engineering) while addressing **conflicting requirements** such as accuracy vs. usability.

From an engineering perspective, this project ensures:

- **Systematic problem-solving** aligned with global engineering education standards.
- **Knowledge integration** of clinical data science and software engineering.

Significant societal impact by supporting early dengue detection in Bangladesh.

Chapter 6

Conclusion

This chapter provides a comprehensive conclusion to the research on dengue fever prediction using machine learning models. It summarizes the major findings of the study, outlines the key limitations that affect generalizability, and highlights potential directions for future work to strengthen the system's accuracy, scalability, and integration into real-world healthcare applications.

6.1 Summary

This study developed and evaluated multiple machine learning models to predict dengue fever using an extensive multi-hospital clinical hematology dataset comprising 2,086 samples collected over approximately six months. Unlike previous systems relying on small or seasonal data, this work leverages a larger, balanced dataset with rich hematology features to improve prediction accuracy and clinical relevance. Ensemble models such as Random Forest and XGBoost achieved the highest performance, demonstrating strong precision, recall, and overall reliability for dengue detection in resource-limited healthcare environments like Bangladesh. Methodologies including hyperparameter tuning and detailed evaluation metrics ensured a thorough assessment of model effectiveness. The study lays a robust groundwork for future integration into clinical workflows to enhance early dengue diagnosis and patient care.

6.2 Limitation

Despite the improved dataset size of 2,086 samples, limitations persist. The focus on hematology data excludes potentially valuable inputs such as environmental, climate, or socio-demographic factors that impact dengue outbreak dynamics. Although the dataset is larger, it is still geographically and temporally limited, which may constrain generalizability to other regions or different time periods. External validation on independent datasets has not been performed, so the model's robustness in varied real-world conditions requires further confirmation. Furthermore, real-time application and integration with healthcare systems remain future goals, which must be addressed for practical implementation and widespread adoption.

This research concludes on a note that it is possible, and provides great future possibilities of using Machine Learning models in enhancing the prediction of Dengue Fever. Implementing these models, especially the Random Forest model, will help in assisting public health authorities to take necessary actions leading to Dengue preventive measures in Bangladesh. Continued model refinement and incorporation of real-time data to improve disease management strategies can ultimately lead to important public health outcomes.

6.3 Future Work

As now proven, the ML-based approach to forecasting dengue has broad scope for new angles of study. An alternative approach to this is the introduction of potentially useful additional parameters, for instance, weather and geographical data giving finer detail on

the micro-ecosystems of Dengue Fever spread. These may increase the precision of the model, since they capture a considerable seasonal and spatial detail.

Like the other directions for future research, these pertain to conducting additional research in other areas, implementing advanced forms of machine learning (i.e. deep learning) capable of extracting intricate information from troves of data that simpler models are unable to comprehend (or, adjusting the hyperparameters of the models in consideration, utilizing Bayesian optimization or genetic algorithms among other things). Furthermore, there are ensemble techniques that tend to perform, or outperform, the rest, and are therefore more optimal for predicting Dengue.

Additionally, timely streaming of data, like live weather information, mosquito population counts, and healthcare reports, can be a challenging and interesting future research direction. They could act faster to outbreaks before any the first cases are even reported, if public health officials had real-time prediction models. On top of that, you can address problems related to your data imbalance using techniques as SMOTE (Synthetic Minority Over-sampling Technique) or cost-sensitive learning, to improve the model performance, and especially when presenting imbalanced datasets. The application to other locations (notably the rural sites or another Southeast Asian country with high-density dengue) will need to expand the geographical footprint of the project to inform a more geographically generalizable model. Moreover, for the models to become user friendly, there will be a necessity to create practical tools (e.g., mobile applications and user-friendly dashboards) that can be easily applied in the daily clinical practice by healthcare professionals with little technical knowledge. In conclusion, this study serves as a proof of concept for the application of machine learning techniques to predict dengue outbreaks at the national scale, but further work is yet required in order to reach acceptable levels for deployment in the real-world and at scale in dengue fever control programming.

References

- [1] Hasan, M. J., Tabassum, T., Sharif, M., Khan, M. A. S., Bipasha, A. R., Basher, A., ... & Amin, M. R., "Comparison of clinical manifestation of dengue fever in Bangladesh: an observation over a decade," *BMC Infect. Dis.*, vol. 21, pp. 1-10, 2021.
- [2] S. Banu, W. Hu, C. Hurst, Y. Guo, M. Z. Islam, and S. Tong, "Space-time clusters of dengue fever in Bangladesh," *Trop. Med. Int. Health*, vol. 17, no. 9, pp. 1086-1091, Sep. 2012.
- [3] S. Sharmin, E. Viennet, K. Glass, and D. Harley, "The emergence of dengue in Bangladesh: epidemiology, challenges and future disease risk," *Trans. R. Soc. Trop. Med. Hyg.*, vol. 109, no. 10, pp. 619-627, Oct. 2015.
- [4] T. Shirin, "Dengue fever in Bangladesh," *Bangladesh Med. Res. Counc. Bull.*, vol. 45, no. 3, pp. 214-215, Dec. 2019.
- [5] M. S. Rahman, A. Karamelic-Muratovic, M. Baghbanzadeh, M. Amrin, S. Zafar, N. N. Rahman, S. U. Shirina, and U. Haque, "Climate change and dengue fever knowledge, attitudes and practices in Bangladesh: A social media-based cross-sectional survey," *Trans. R. Soc. Trop. Med. Hyg.*, vol. 115, no. 1, pp. 85-93, Jan. 2021.
- [6] M. M. Rahman, A. R. Islam, S. J. Khan, K. N. Tanni, T. Roy, M. R. Islam, M. A. Rumi, M. S. Sakib, M. A. Quader, N. U. Bhuiyan, and M. A. Chisty, "Dengue fever responses in Dhaka City, Bangladesh: a cross-sectional survey," *Int. J. Public Health*, vol. 67, p. 1604809, Aug. 2022.
- [7] K. M. Rahman, Y. Sharker, R. A. Rumi, M. U. Khan, M. S. Shomik, M. W. Rahman, S. M. Billah, M. Rahman, P. K. Streatfield, D. Harley, and S. P. Luby, "An association between rainy days with clinical dengue fever in Dhaka, Bangladesh: findings from a hospital based study," *Int. J. Environ. Res. Public Health*, vol. 17, no. 24, p. 9506, Dec. 2020.
- [8] M. Z. Islam, S. Rutherford, D. Phung, M. N. Uzzaman, S. Baum, M. M. Huda, M. A. Asaduzzaman, M. R. R. Talukder, and C. Chu, "Correlates of climate variability and dengue fever in two metropolitan cities in Bangladesh," *Cureus*, vol. 10, no. 10, 2018.
- [9] M. A. Mamun, J. Misti, M. D. Griffiths, and D. Gozal, "The dengue epidemic in Bangladesh: risk factors and actionable items," *Lancet*, vol. 394, no. 10215, pp. 2149-2150, Dec. 14, 2019.
- [10] M. Rahman, K. Rahman, A. K. Siddique, S. Shoma, A. H. M. Kamal, K. S. Ali, A. Nisaluk, and R. F. Breiman, "First outbreak of dengue hemorrhagic fever, Bangladesh," *Emerg. Infect. Dis.*, vol. 8, no. 7, p. 738, 2002.
- [11] P. Mutsuddy, S. T. Jhora, A. K. M. Shamsuzzaman, S. M. G. Kaisar, and M. N. A. Khan, "Dengue situation in Bangladesh: an epidemiological shift in terms of morbidity and mortality," *Can. J. Infect. Dis. Med. Microbiol.*, vol. 2019, p. 3516284, 2019.
- [12] M. R. Mobarak, M. R. Islam, A. T. Bhuiyan, N. Akand, and F. Begum, "Evaluation of dengue fever in a tertiary care children hospital of Bangladesh," *Northern Int. Med. Coll. J.*, vol. 9, no. 1, pp. 274-277, 2017.
- [13] A. A. Akram, "Alarming turn of dengue fever in Dhaka city in 2019," *Bangladesh J. Infect. Dis.*, vol. 6, no. 1, pp. 1-2, Aug. 17, 2019.
- [14] A. S. Alam, S. A. Sadat, Z. Swapan, A. U. Ahmed, M. N. Karim, H. K. Paul, and S. Zaman, "Clinical profile of dengue fever in children," *Bangladesh J. Child Health*, vol. 33, no. 2, pp. 55-58, 2009.

- [15] M. Miah, M. B. Hossain, S. N. Jannat, M. R. Karim, M. R. Rahman, Y. Arafat, and F. H. Pingki, "Assessing the impact of climatic factors on dengue fever transmission in Bangladesh," *Aerobiologia*, vol. 40, no. 2, pp. 233-245, Jun. 2024.
- [16] [16] I. Sarker, M. R. Karim, S. E-Barket, and M. Hasan, "Dengue fever mapping in Bangladesh: A spatial modeling approach," *Health Sci. Rep.*, vol. 7, no. 6, p. e2154, 2024.
- [17] R. C. LaRocque, R. F. Breiman, M. D. Ari, R. E. Morey, F. A. Janan, J. M. Hayes, M. A. Hossain, W. A. Brooks, and P. N. Levett, "Leptospirosis during dengue outbreak, Bangladesh," *Emerg. Infect. Dis.*, vol. 11, no. 5, p. 76, 2005.

213-15-4371

ORIGINALITY REPORT

19%	13%	13%	13%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	4%
2	Submitted to BB9.1 PROD Student Paper	1%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
4	www.mdpi.com Internet Source	1%
5	ijream.org Internet Source	1%
6	Submitted to Salem State University Student Paper	<1%
7	Submitted to Dublin Business School Student Paper	<1%
8	Submitted to University of New South Wales Student Paper	<1%
9	Submitted to Coventry University Student Paper	<1%
10	Submitted to Nottingham Trent University Student Paper	<1%
11	pnrjournal.com Internet Source	<1%
12	www.coursehero.com Internet Source	<1%

www.saudijournals.com

