Dengue detection using Machine Learning Algorithms

This Thesis is Submitted in Fulfillment
of the Requirements for the Degree of

Bachelor of Science (B.Sc.)

in

Computer Science and Engineering (CSE)

by

Md. Reyad Hossain (C191083)
Md. Hasibul Hossain (C191093)
Nazmuz sakib Turhan (C191107)

TO
FACULTY OF SCIENCE AND ENGINEERING
INTERNATIONAL ISLAMIC UNIVERSITY CHITTAGONG
Spring 22

Md.HasibulHossain

DECLARATION

We hereby affirm the following statements regarding our thesis:

1. The thesis has been successfully completed as part of our undergraduate degree program at International Islamic University Chittagong.
2. The thesis work does not contain any previously published or third-party content without proper citation.
3. The thesis work has not been previously submitted for any other degree or diploma at any other university or institution.
4. We have appropriately acknowledged all significant sources of contribution in the thesis.

Student's Full Name and Metric ID:

Md. Reyad Hossain **(C191083)**

Md. Hasibul Hossain **(C191093)**

Nazmuz Sakib Turhan **(C191107)**

SUPERVISOR'S DECLARATION

I formally state that I have examined this thesis and claim it to be of sufficient quality and scope to be granted for the undergraduate degree of Bachelor of Science in Computer Science and Engineering.

Dr. Shahidul Islam Khan
Associate Professor
Department of Computer Science and Engineering
International Islamic University Chittagong

DEDICATION

This thesis report is dedicated to us, our supervisor, and our family. The teamwork was satisfactory and the family's support was incredibly amazing. Our dedicated and hard-working Supervisor, Dr. Shahidul Islam Khan, has been a constant support throughout these months. In this document, the contributions are acknowledged too.

ACKNOWLEDGMENT

To start with, All the praises to the Almighty Allah, for his mercy because of whom we were able to finish our thesis despite having so many obstacles. Secondly, we would like to thank our Supervisor, Dr. Shahidul Islam Khan, for his continuous effort and guidelines from the beginning of our research.

ETHICAL STATEMENT

Hereby we state that, none of the unethical practices were used in the completion of our thesis work. The data we used for the research purpose are original. We carefully checked every citation we used here. The three writers of the work accept all the liabilities forany kind of violation of the thesis rule.

ABSTRACT

The dengue virus, a global health concern with an estimated 400 million annual infections, has posed a critical challenge worldwide. Between January 1 and August 7, 2023, Bangladesh experienced a concerning surge in dengue cases, reporting over 69,000 confirmed cases and 327 deaths according to the Ministry of Health and Family Welfare. Timely and accurate diagnosis is necessary for intervention and treatment optimization, possibly relieving severe complications and saving lives. In response, this paper introduces a machine-learning approach to predict outbreaks in Chittagong and Cox's Bazar. Through careful analysis of region-specific survey data, we identified local triggers for dengue spikes. Our data-driven model empowers healthcare officials with predictive capabilities to address future outbreaks, safeguarding communities. The study featuresa unique real-time dataset collected from healthcare institutions across Chittagong and Cox's Bazar, involving collaboration with esteemed public and private institutions. By using diverse data, the dataset aims to unveil hidden insights into dengue outbreaks, guiding accurate predictions and effective prevention measures. The comprehensive patient dataset, including diagnoses, medical history, and symptoms, underwent precise model training with a 70:30 split. we have applied various machine learning algorithms, namely - SVM, Decision Tree, XGBoost, Naive Bayes, Random Forest, K-NN, Logistic Regression, and LDA. The Random Forest demonstrated accuracy of 98.92 percent on both training and test data. Performance assessment included a confusion matrixand macro averages for precision, recall, and F1-measure scores.

Keywords: SVM, Decision Tree, XGBoost, Naive Bayes, Random Forest, K-NN, Lo- gistic Regression, LDA

Table of Contents

List of Figures

List of Tables

List of Algorithms

ABBREVIATION

The following list provides descriptions of various symbols and abbreviations that will be utilized in the subsequent sections of the document.

SVM: Support Vector Machine
DT: Decision Tree
XGBoost: Extreme Gradient Boosting
NB: Naive Bayes
RF: Random Forest
K-NN: K-Nearest Neighbors
LR: Logistic Regression
LDA: Linear Discriminant Analysis

Chapter 1

Introduction

1.1 Overview

There are numerous conventional methods for DENV testing. We have used NS1 based antigen testing using machine learning algorithms in our proposed model. This model utilizes a 20-feature architecture on raw dataset of 308 patient records from renowned institutions and after going through several stages brought to a form suitable for precise diagnosis. Ultimately, if the model demonstrates sufficient accuracy, it could evolve into a valuable tool for both patients and doctors, guiding treatment decisions and contributing to more effective dengue prevention and control strategies.

1.2 Motivation and Scope of the Research

The total count of individuals afflicted with the illness known as dengue in Bangladesh within the current year is approximately two times greater than the collective number of cases documented during the preceding 19-year period. According to the DG of Health Services, no. of people afflicted with Dengue fever in Bangladesh from 2000 to 2022 amounted to 2,43,748. The number of deaths caused by dengue fever reached a total of 1,705 in the year 2023.

1.3 Problem Statement

Dengue is the most death-defying or perilous disease in our country. People suffering from these types of diseases become depressed for not recovering fully. It becomes more concern when the disease is Dengue fever. Dengue patients may have a higher chance ofdeath. Sometimes long-time treatments are required to take which are very costly for a developing country like Bangladesh. Thus, most people often cannot continue medical diagnosis resulting in severe health conditions.

1.4 Contribution of the thesis

The main contributions of this thesis are as follows:

1. We have collected raw 308 patients' data comprising of dengue and non-dengue patients from renowned medical institutions and then processed them for making them diagnosis ready.
2. We have performed NS1 based antigen testing using machine learning algorithms which would help in early detection, real-time monitoring and enhanced decision support of the disease etc.
3. We have demonstrated our model through developing a web-based application that utilizes ML algorithms to predict and diagnose Dengue cases, providing users with a user-friendly interface.

1.5 Organization of the thesis

In Section 1 of the report, a concise and clear overview of the research objectives and their significance is provided. Section 2 focuses on the comprehensive literature review, presenting the present form of research in the area.

Chapter 2

Literature Review

2.1 Introduction

Recent work, objectives, and enthusiasm for our work have been analyzed. Now we will describe and discuss in this section the previous studies related to our work. The previous work-study is necessary to get the knowledge and ideas. All of these researchworks are already accomplished by using different kinds of data mining techniques. For example, SVM, Decision Tree, Naive Bayes, Random Forest, K-NN, Logistic Regression, and so on have been used. In these several works, except a few works, no research worked based on physical, clinical, and diagnosis data.

2.2 Dengue-related research work

2.2.1 Work done by implementing SVM:

Rajeev Kapoor et al [1] used the SVM classification model after pre-processing thedataset in their model. Found four important factors in detecting dengue at the very beginning. Implementing more parameters and more data can increase the accuracy in the future.

Samina Amin et al [2] tracked the movement of seasonal epidemics and analyzed Twitter conversations to estimate epidemiological data using Social Media analysis. Proposed machine learning-based approach to detect dengue and flu outbreaks in social media platform Twitter using four machine learning algorithms: RF, KNN, SVM and DT, of which RF classifier outperformed them. The collection of more data can give the best accuracy in the future.

V. Janani et al. [3] used the diverse strengths of four distinct algorithms – Multilayer Perceptron, Support Vector Machine, Sequential Mining Optimization, and WEKA in their model. The impressive predictive accuracy achieved by the SMO algorithm vali- dates the chosen parameters as trustworthy markers for dengue diagnosis. Implementing more parameters and more data can increase the accuracy in the future.

NurulAzam Mohd Salim et al [4] imported their dataset into IBM SPSS Modeler, built various predictive models, and assessing of their performance was done using the analysis node. Of all the Ml algorithms applied, SVM outperformed others by showing 70 percent accuracy. Future research on dengue prediction models will explore the potential of bio-inspired algorithms.

2.2.2 Work done by implementing Decision Tree:

Dhiman Sarma et al. [5] utilized a large patient dataset, applying thorough preprocess-ing techniques and advanced feature engineering. The decision tree algorithm achieved an impressive 79 percent accuracy in classifying three types of dengue fever. New infor- mation beyond known features and risk factors for dengue fever will be explored in the future.

Song Quan Ong et al. [6] used a mix of seven diverse machine learning algorithms, each with unique strengths for prediction. Compared to traditional approaches like logistic regression, Naive Bayes, Decision tree, and SVM, the ensemble model showed excellent performance on all evaluation criteria. Future research will focus on identifying the most informative features, exploring adaptable model architectures, and expanding the data landscape.

2.2.3 Work done by implementing logistic regression:

T.Sajana et al [7] implemented MLP, C-4.5, regression tree for accurate diagnosing in their model. More accurate and timely results can be obtained by using machine learning algorithms precisely. Using different models and collecting more data can help to detect dengue more efficiently in the future.

Jun Kit Chaw et al. [8] used machine learning, prediction, and ensemble learning to hold immense potential for real-world applications. Machine learning algorithms namely - Logistic regression, DT, SVM, neural networks, and ensemble learning of bagging and boosting are implemented, of them, the bagging algorithm outperforms others with a 14.5 percent improvement from the individual decision tree. Driven by the pursuit of

innovative solutions, future research will focus on expanding the evaluation and adoption of machine-learning models for dengue prediction promises significant advantages.

S. Chattopadhyay et al. [9] drew upon a diverse arsenal of statistical machine-learning techniques. Driven by the pursuit of improved healthcare outcomes, this study showcases a pioneering approach to diagnosis that merges the best of clinical expertise with data-driven analysis. MLR and MnLR algorithms are used, of which MnLR outperformed. With a keen eye on enhancing both accuracy and generalizability, future efforts will center on extending the scope of the present SML classifier.

2.2.4 Work done by implementing xgboost:

Yiran E. Liu et al. [10] integrated 11 datasets, conducted multi-cohort analysis, trained an XGBoost model, measured DEG expression, and compared accuracy with clinical signs for predicting SD. A smart 8-gene XGBoost model accurately predicted SD pro- gression in a large, independent group using a variety of public data. Future research must bridge the gap between laboratory achievements and tangible patient benefits to maximize the positive impact of the 8-gene model.

Donald Salami et al. [11] identified the optimal model, capable of predicting future importation events and informing proactive public health strategies by training and evaluating four distinct classifiers. Several machine learning algorithms namely - pls, glmnet, RF and xgboost were implemented. Of them, RF and xgboost outperformed the others with 97 percent accuracy. Future research will focus on Prioritizing practicality and interpretability for public health applications, and prepared data by aggregating yearly dengue incidence rates from source countries.

2.2.5 Work done by implementing Random Forest:

Permatasari Silitonga et al. [12] analyzed lab characteristics and data from different observation days. The RF classifier, utilizing a 10-fold cross-validation scheme, achieved the highest 58 percent accuracy compared to the other classifiers. Future research may involve developing a new ANN architecture for the same data to achieve a higher accuracy model.

Sheng-Wen Huang et al. [13] paved the way for more accurate and impactful prognostic models across various domains. Various machine learning algorithms have been used, namely - LR, RF, GBM, SVM and ANN. Of them, ANN showed the highest average discrimination area under the balance accuracy. Dedicated to the pursuit of robust

and reliable solutions, the developers of this model acknowledge the need for further validation with external cohorts in future studies.

2.2.6 Work done by implementing other algorithms:

Satya Ganesh Kakarla et al. [14] embarked on a methodological journey, employing a unique blend of machine learning (SVR, GBM), deep learning (LSTM), and statisti- cal methods (VAR). The LSTM model outperformed other models by giving the best accuracy in overall dengue case prediction. Future research will focus on a Proactive approach like Prediction that optimizes resource utilization, ensures readiness for crises,and ultimately improves population health outcomes.

Elisa Mussumeci, Fl´avioCodeco Coelho [15] employed in this study to predict dengue incidence beyond the scope of the training data an LSTM recurrent neural network model. Beyond the technical prowess of algorithms, the potential of machine learning models to forecast dengue incidence time series holds immense societal value. Future research will focus on expanding the evaluation and adoption of machine-learning models for dengue prediction promises significant advantages.

ANNALISA APPICE et al. [16] developed a powerful tool for predicting dengue by exploring established predictive methods in a new multi-stage machine learning framework called ”AutoTiC-NN”. A trend association-based nearest neighbour predictor performed the prediction. Studying dengue transmission properties in association with various variables and extending AutoTiC- NN will be done in the future.

S.Saturi [17] forecasted models and K- means clustering method to develop a forecast model. This research unearthed the fundamental components that hold the key to accurate dengue forecasting. This research sheds light on the critical vulnerabilities of current dengue forecasting models, guiding future research toward targeted solutions.

Felestin Yavari Nejad, Kasturi Dewi Varathan [18] unveiled the hidden language of climate data, we performed detailed correlation analyses to decode the climatic fac- tors significantly influencing dengue outbreaks. The Bayes Network model embraced a newly identified meteorological risk factor, boosting its accuracy in predicting dengue outbreaks to a remarkable 92.35 percent. By embracing the wisdom of a previously obscure meteorological factor, the Bayes Network model, a sophisticated probabilistic framework, achieved a previously unimaginable 92.35 percent accuracy in predicting dengue outbreaks.

Abrar Noor Akramin Kamarudin et al. . [19] embraced the potential of machine learning algorithms, this research proposed a new conceptual framework designed to optimize MBD outbreak prediction accuracy. The enhanced MBD outbreak prediction framework, bolstered by the inclusion of the entomological index, holds immense potential for early warning systems. The proposed framework should focus on tackling the challenge of data access head-on, offering a novel solution for MBD outbreak prediction in the future.

Aima Aziz, Azka Aziz [20] employed both prediction and forecasting techniques within the machine learning framework and enriched the training process by incorporating historical data with simulated projections. Building upon the foundation laid by this research, future studies that delve deeper into the temporal dynamics of these results have the potential to revolutionize clinical practice.

2.3 Summary

We detected that except few most of them did not combine medical history data and clinical and personal data of patients for the classification. In our work, we tried to focus on the field of ns1-based antigen testing using machine learning algorithms for early detection of dengue. The discussed research papers' authors have mentioned that they have used publicly available data from databases, or secondary data. However, we are using real-time data that are not available directly from databases or records from the healthcare industry.

Chapter 3

Methodology

3.1 Introduction

We have investigated different types of procedures. We have tried to go through different models from past research. By examining different algorithms, we tried to find the appropriate result to gain better accuracy.

3.2 Our Approach:

The phases of our approach are described below-

3.2.1 Data Collection Process

We have collected a total of 308 patients' data with 20 attributes from renowned public and private hospitals of Chittagong and Cox's Bazar. In the case of public hospitals, different types of patients visit there regularly rather than in private hospitals.

We can divide our survey data attribute into three parts.

1. Personal information of the patients.
2. Clinical information
3. Test and treatment values.

Personal information of a patient includes:

1. Age: Age of a patient.
2. Sex: Gender of a patient.
3. Work: This attribute has been selected to know his daily working activity.

Clinical Data of a patient includes:

1. Headache.
2. Muscle Pain.
3. Bleeding.
4. Skin rash.
5. Vomiting.
6. Diarrhea.
7. Fatigue.
8. Rapid Breathing.

Test and treatment-related attributes include:

1. Systolic Blood Pressure.
2. Diastolic Blood Pressure.
3. Cough.
4. Pulse.
5. Low level of platelet.
6. IgG.
7. IgM.
8. NS1.
9. Fever.

3.2.2 Data Preprocessing and Transformation

Real-world data tend to be noisy, missed or incomplete or may be inconsistent. Preprocessing makes data understandable by various processes and solves those issues. We have used mean imputation to replace our missing values. Data preprocessing of our dataset includes:

3.2.2.1 Normalization

We performed normalization and converted values of our initial dataset into 0-2 values depending on attributes. We are grateful to Doctor for assisting in the normalization of our dataset.

3.2.2.2 Handling Missing values

We had missing values of LLBplatelet = 7, Bleeding = 11, and vomiting = 13. We have used mean value imputation for handling these missing values.

3.2.3 Dataset splitting

After preprocessing, we obtained pure and error-free data and split our dataset into different ratios of train and test dataset whereas we get the best performance with the ratio of 70 percent training and 30 percent testing dataset.

3.2.4 Model Training

Let's take a closer look at each model's architecture, crucial information, and functionality:

Algorithm 1 SVM:

1: Creates a hyperplane in high-dimensional space to separate classes by maximizing the difference between them. The support vectors, or data points nearest to the hyperplane, form the decision boundary.

2: Capable of handling huge feature spaces efficiently. Effective in circumstances where the number of dimensions outnumbers the number of samples.

3: A kernel function is used to map input data to a higher-dimensional space. Maximize the margin to find the hyperplane that best separates the classes.

Algorithm 2: DT Classifier:

1: Represents a tree-like structure, with each internal node representing a feature, each branch indicating a decision based on that feature, and each leaf node representing a class label. Splits the dataset into subsets depending on feature values to reduce impurity.

2: Can work with both numerical and category data. Overfitting is possible, particu- larly with deep trees.

3: The feature space is recursively partitioned into smaller portions by selecting the best feature and threshold for each node.

Algorithm 3: XGBoost:

1: The ensemble learning method combines the predictions of numerous decision trees generated successively.

2: Gradient boosting is used to optimize the model by minimizing a loss function. To avoid overfitting, regularization techniques are implemented.

3: Creates a sequence of decision trees, with each succeeding tree attempting to rem- edy the flaws of the prior one. Combines individual tree forecasts to generate the outcome.

Algorithm 4: GNB:

1: A probabilistic classifier based on Bayes' theorem with the premise of feature inde- pendence. Given the input features, this function calculates the likelihood of each class.

2: It only takes a modest quantity of training data to estimate parameters efficiently. Performs well when presented with irrelevant features.

3: Using the joint probability distribution of the characteristics and class labels, the likelihood of each class is estimated.

Algorithm 5: RF:

1: During training, the ensemble learning method generates numerous decision trees and outputs the mode of the classes or the mean prediction of each tree.

2: Overfitting is reduced in comparison to individual decision trees. It includes built-in feature importance scores.

3: Creates a large number of decision trees and then averages or votes on their forecasts.

Algorithm 6: KNN:

1: A non-parametric, instance-based learning technique that classifies instances using the majority vote of their k nearest neighbors.

2: The algorithm is simple and intuitive. Large datasets can result in slow prediction times.

3: By comparing it to its k nearest neighbors in the training data and assigning the majority class label among those neighbors, a new instance is classified.

Algorithm 7: LR:

1: Linear binary classification model that employs a logistic function to predict the probability of an event occurring.

2: The model is simple and easy to understand. Produces probability estimates for class membership.

3: A logistic function is used to model the probability of the binary result, mapping the input attributes to probabilities ranging from 0 to 1.

Algorithm 8: LDA:

1: It is a linear classification algorithm that determines the optimal linear combination of features for distinguishing two or more classes.

2: Assumes that the features across every category are evenly distributed and that the matrices for covariance are the same across all classes. Can be used to minimize dimensionality.

3: Projects the data into a lower-dimensional space while maximizing the distance between classes and minimizing variance within each class.

The above descriptions provide a thorough understanding of the individual model's characteristics and inner workings.

3.3 Workflow Diagram:

The research methodology of our approach is below:

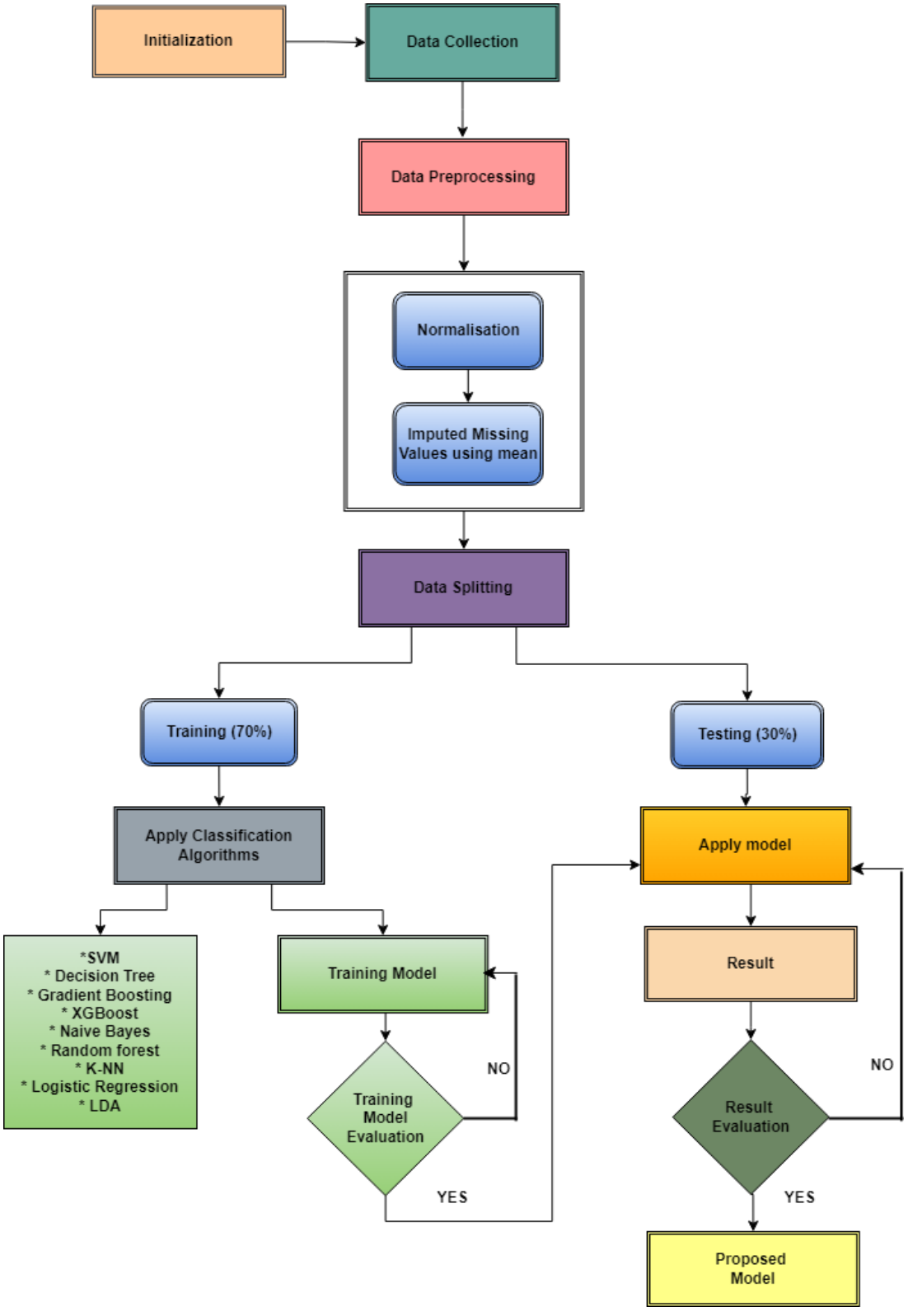


Fig. 3.1. Schematic diagram of Dengue prediction model

3.4 Architecture of Random Forest:

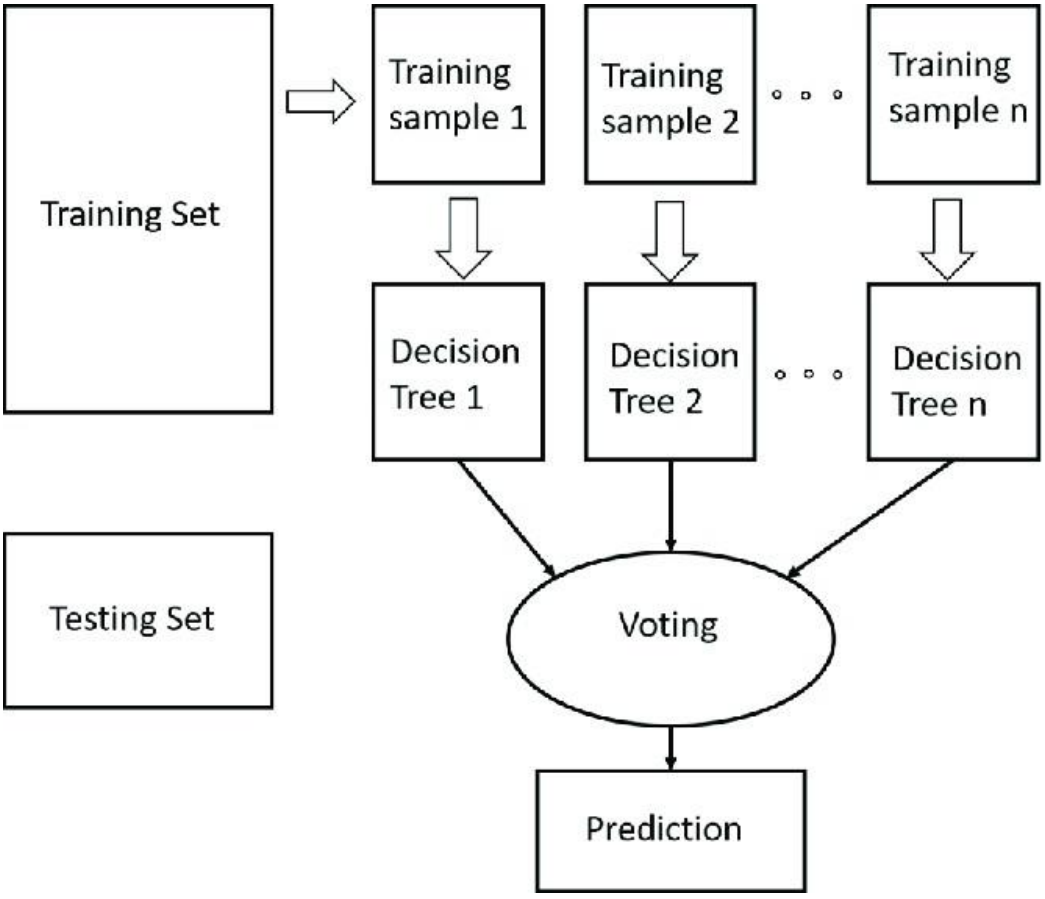


Fig. 3.2. The architecture of the Random Forest Classifier [21]

3.5 Summary:

We have gone through different procedures in the overall methodology part, namely: Data collection, Data Preprocessing, Dataset splitting, and then Model training. In the case of all the algorithms discussed above, Random Forest outperformed them. In the next chapter, we will show results and discussions related to detecting dengue cases.

Chapter 4

Results and Discussions

4.1 Introduction

The experiment aimed to employ machine learning techniques namely - SVM, Decision Tree, XGBoost, Naive Bayes, Random Forest, K-NN, Logistic Regression, and LDA toidentify dengue cases. The primary goal is to investigate the usefulness of several algo- rithms in reliably detecting and forecasting dengue cases, using a wide set of methodolo-gies. In this chapter, we will discuss and show the implementation results with related algorithms and Tools.

4.2 Dataset and Experimental Settings

We have collected the dataset, pre-processed it, split it into train and test, and thus made it ready to train our model.

4.2.1 Experimental Tool and Environment

For the experiment Laptop, internet connection, and Google Drive access were necessary.

4.2.2 Programming Language

Our whole work was performed by using Python version 3.7, which was compatible with any virtual environment. Python is easy to implement and highly effective for research purposes. All the necessary libraries were available in Python, making the process easier.

4.2.3 Integrated Development Environment (IDE)

Google Collaboratory: Google Collaboratory: It is a free cloud service that Google released to make machine learning and research easier. It provides great runtime as well as regular runtime for shaping. The experiment was then run on a laptop.

4.3 Libraries

NumPy: It is a library for Python programming language. It is used for multi-dimensional arrays and metrics.

Pandas: It is a Python toolkit for rapid data processing and analysis, containing struc- tures such as DataFrame and Series for labeled datasets.

Scikit-learn: Scikit-learn is a popular open-source machine-learning library for Python. It provides simple and efficient tools for data analysis and modeling, including various machine-learning algorithms for classification, regression, clustering, dimensionality re-duction, and more.

Matplotlib: Matplotlib is an extension of NumPy for numerical mathematics. It is used for creating plots and visualizations during data analysis.

4.4 Dataset

The below figure shows the attribute table of our patients' dataset:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Age	Sex	Work	BpSys	Bpdias	Pulse	Cough	Fever	LLBplatele	Headache	Muscle pa	Bleeding	Skinrash	Vomiting	Diarrhoea	IgG	IgM	Fatigue	Rapid bre	NS1
2	32	1	2	120	80	83	0	0	0	0	1	0	0	0	0	0	0	1	0	0
3	28	1	1	110	80	70	1	1	0	1	1	0	0	1	1	0	0	1	0	0
4	30	1	1	110	70	75	0	0	0	0	1	0	0	0	0	0	0	0	0	0
5	2	1	0	110	80	90	0	1	0	0	0	0	0	1	0	0	0	1	1	0
6	22	1	1	140	100	77	0	0	0	1	1	0	0	1	1	0	0	1	1	0
7	34	1	2	130	100	78	0	0	0	0	1	0	0	0	1	0	0	1	1	0
8	45	0	0	110	70	90	1	2	0	1	1	0	0	1	0	0	0	1	1	0
9	34	0	1	90	60	88	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	28	0	0	120	70	90	1	2	0	0	1	0	0	0	0	0	0	0	0	0
11	25	0	1	90	60	82	0	1	0	0	0		0	0	0	0	0	1	1	0
12	18	1	0	80	50	81	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	22	0	1	85	60	74	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	37	1	1	70	60	70	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	17	1	0	90	60	88	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	29	0	1	80	60	71	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	45	0	0	110	70	108	1	2	0	0	0	0	0	0	0	0	0	0	0	0
18	34	1	1	130	90	85	1	2	0	1	1	0	0	1	0	0	0	0	0	0
19	26	1	0	110	70	87	1	2	0	1	1	0	0	1	0	0	0	0	0	0
20	23	0	0	90	60	61	0	0	0	1	0	0	0	1	0	0	0	1	1	0
21	32	1	1	90	60	79	0	0	0	0	0	0	0	1	0	0	0	0	0	0

Fig. 4.1. Attributes table

4.5 Data preprocessing:

Data preprocessing is essential for converting raw datasets into pure and error-free datasets suitable for analysis.

A figure demonstrating different stages of Data Pre-Processing is shown below:

Fig. 4.2. DATA PREPROCESSING

4.5.1 Normalized dataset:

Attributes name, their descriptions, and their possible values have been shown in the figure below:

Sl. No.	Variables	Descriptions	Possible Values (Scores)
1	Age	Age of patients	Different numerical values
2	Sex	Gender of patients	0 = Female, 1 = Male
3	Work	Activity of occupation	0 = Sedentary (less active), 1 = Moderately active, 2 = Active
4	Cough	Body Temperature	0 = No, 1 = Yes
5	BpSys	Blood pressure systolic	Different numerical values
6	BpDias	Blood pressure diastolic	Different numerical values
7	Pulse	Pulse rate	Different numerical values
8	Cough	Cough/Squeezing	0 = No, 1 = Yes
9	Fever	Temperature	High = 2, medium = 1, normal = 0
10	HbLevel	Level of platelet	High = 2, medium = 1, normal = 0
11	Headache	Headache	True = 1, False = 0
12	Muscle pain	Pain in body	True = 1, False = 0
13	Bleeding	Bleeding with stool	True = 1, False = 0
14	Skinrash	Skin rash in body	True = 1, False = 0
15	Vomiting	Vomiting frequency	True = 1, False = 0
16	Diarrhoea	Diarrhoea frequency	True = 1, False = 0
17	IgG	Test of IgG	High = 1, normal = 0
18	IgM	Test of IgM	High = 1, normal = 0
19	Fatigue	Tiredness or weakness	True = 1, False = 0
20	Rapid breathing	Rapidity of respiration	True = 1, False = 0
21	NS1	Test of NS1	Positive = 1, Negative = 0

TABLE 4.1: Summary of variables and their possible values.

In this data set, the columns are filled according to the Score of the previously described dataset variable table.

4.5.1.1 Handling Missing Values

Figures below show before and after missing data handling scenarios in counting and heatmap form :

```
Age                0
 Sex               0
Work               0
BpSys              0
Bpdias             0
Pulse              0
Cough              0
Fever              0
LLBplatelet        7
Headache           0
Muscle pain        0
Bleeding          11
Skinrash           0
Vomiting          13
Diarrhoea          0
IgG                0
IgM                0
Fatigue            0
Rapid breathing    0
NS1                0
dtype: int64
```

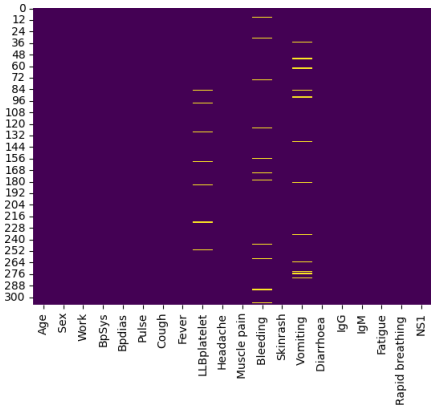


Fig. 4.3. Before missing data handling

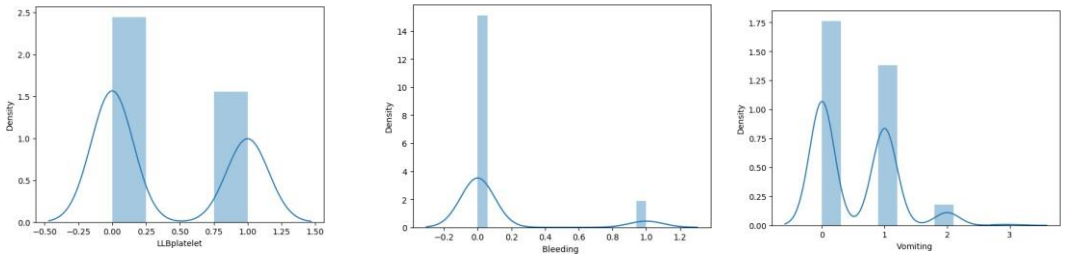


Fig. 4.4. Before missing data handling of LLLBplatelet, Bleeding and Vomiting

```
Age                0
 Sex               0
Work               0
BpSys              0
Bpdias             0
Pulse              0
Cough              0
Fever              0
LLBplatelet        0
Headache           0
Muscle pain        0
Bleeding           0
Skinrash           0
Vomiting           0
Diarrhoea          0
IgG                0
IgM                0
Fatigue            0
Rapid breathing    0
NS1                0
dtype: int64
```

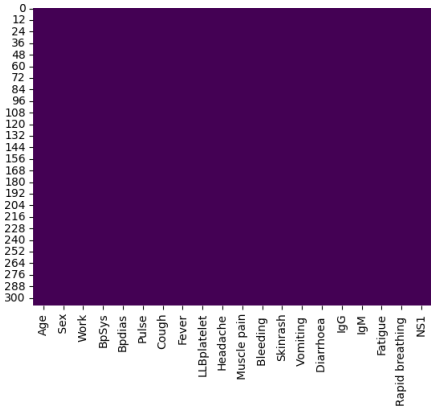
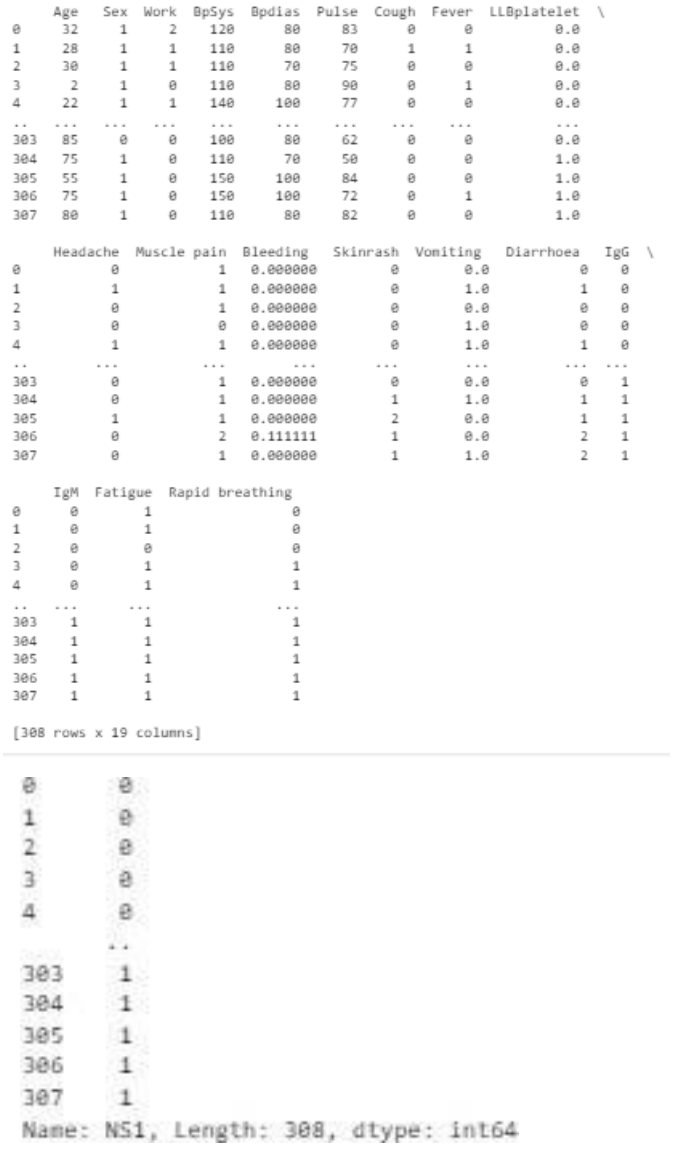


Fig. 4.5. After missing data handling

4.5.1.2 Dataset Splitting

The below figures show our dataset splitting into test and train data :



```
     Age   Sex  Work  BpSys  Bpdias  Pulse  Cough  Fever  LLBplatelet  \
0     32     1     2    120      80     83      0      0          0.0
1     28     1     1    110      80     70      1      1          0.0
2     30     1     1    110      70     75      0      0          0.0
3      2     1     0    110      80     90      0      1          0.0
4     22     1     1    140     100     77      0      0          0.0
..   ...   ...   ...    ...     ...    ...    ...    ...          ...
303   85     0     0    100      80     62      0      0          0.0
304   75     1     0    110      70     50      0      0          1.0
305   55     1     0    150     100     84      0      0          1.0
306   75     1     0    150     100     72      0      1          1.0
307   80     1     0    110      80     82      0      0          1.0

     Headache  Muscle pain  Bleeding   Skinrash  Vomiting   Diarrhoea   IgG  \
0           0            1   0.000000         0       0.0           0    0
1           1            1   0.000000         0       1.0           1    0
2           0            1   0.000000         0       0.0           0    0
3           0            0   0.000000         0       1.0           0    0
4           1            1   0.000000         0       1.0           1    0
..        ...          ...        ...       ...       ...         ...  ...
303         0            1   0.000000         0       0.0           0    1
304         0            1   0.000000         1       1.0           1    1
305         1            1   0.000000         2       0.0           1    1
306         0            2   0.111111         1       0.0           2    1
307         0            1   0.000000         1       1.0           2    1

     IgM  Fatigue  Rapid breathing
0      0        1                0
1      0        1                0
2      0        0                0
3      0        1                1
4      0        1                1
..   ...      ...              ...
303    1        1                1
304    1        1                1
305    1        1                1
306    1        1                1
307    1        1                1

[308 rows x 19 columns]
```

```
0      0
1      0
2      0
3      0
4      0
      ..
303    1
304    1
305    1
306    1
307    1
Name: NS1, Length: 308, dtype: int64
```

Fig. 4.6. Dataset splitting

4.5.1.3 Data Visualization

Below is the distribution of results:

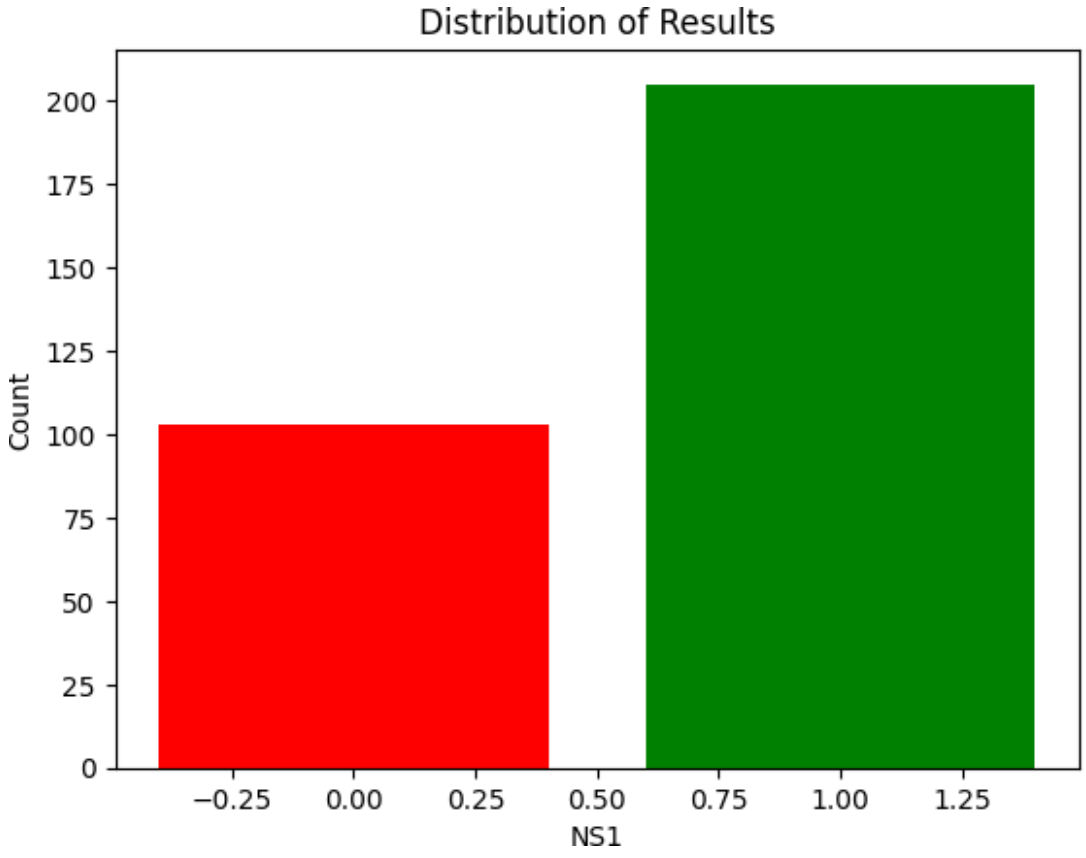


Fig. 4.7. Distribution of results

Below is a correlation heatmap of dengue detection attributes:

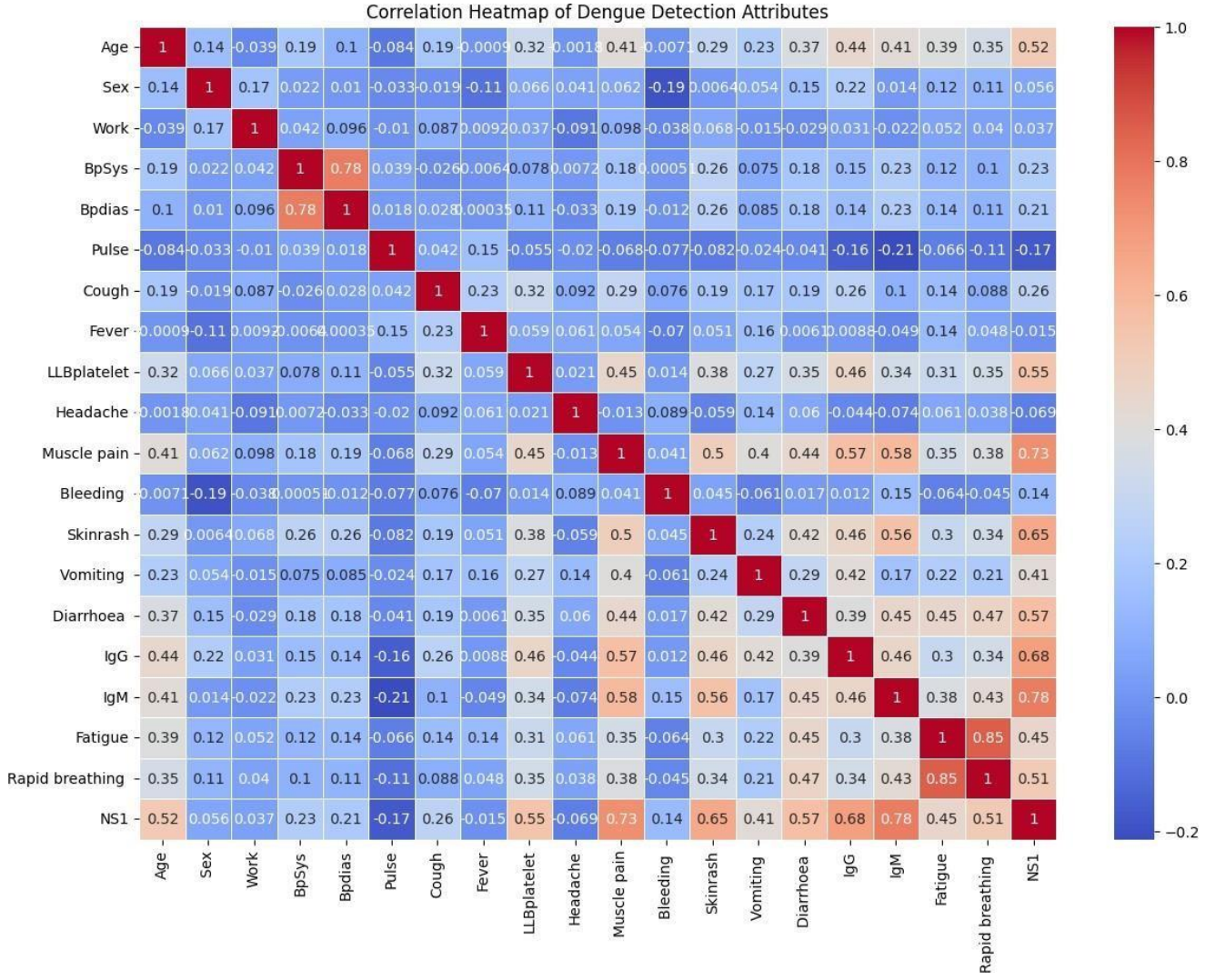


Fig. 4.8. Correlation heatmap of dengue detection attributes

Below is the Count plot for the Categorical Binary values:

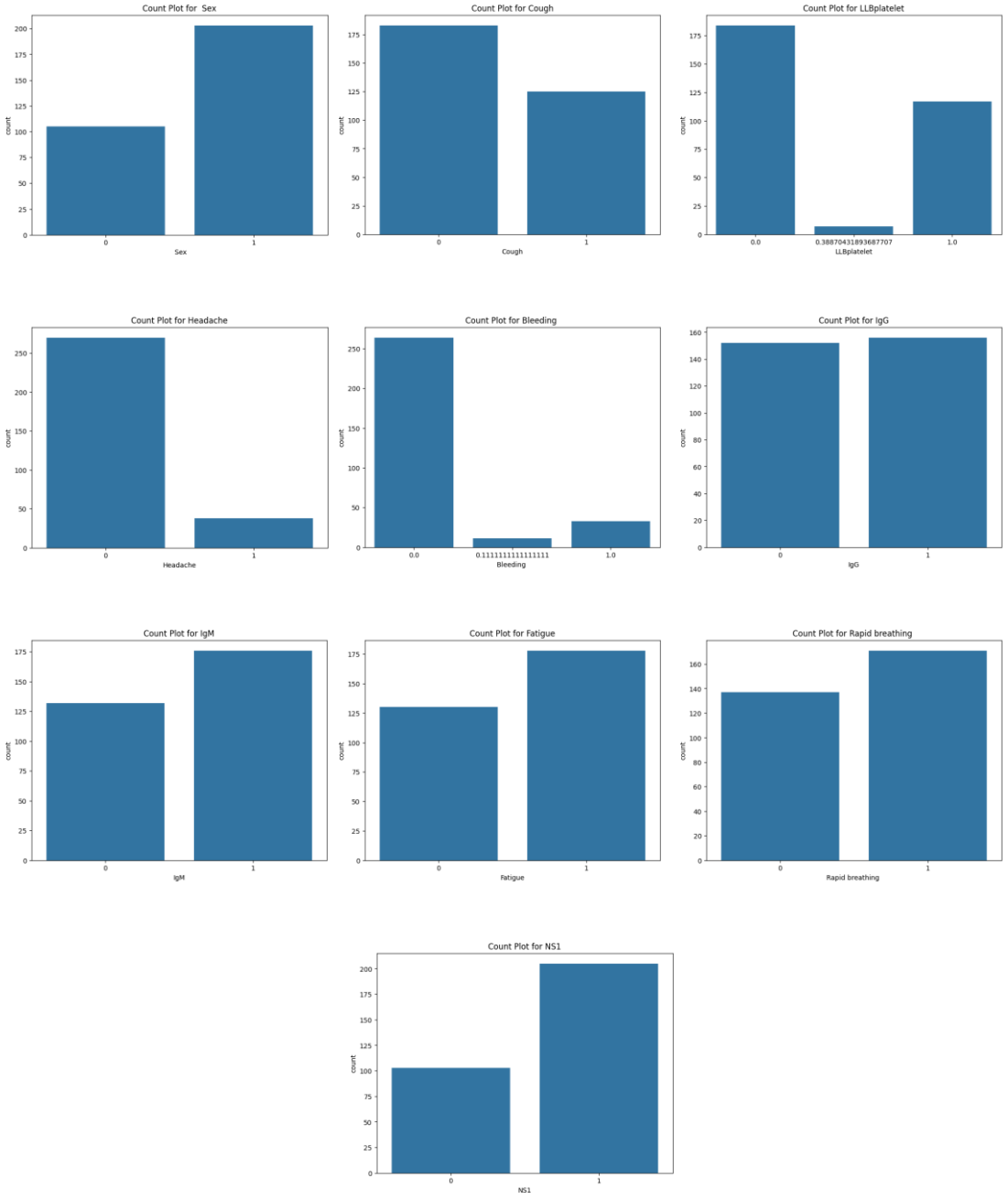


Fig. 4.9. Count plot for the Categorical Binary values

Below is the Box plot for the Categorical Non Binary values:

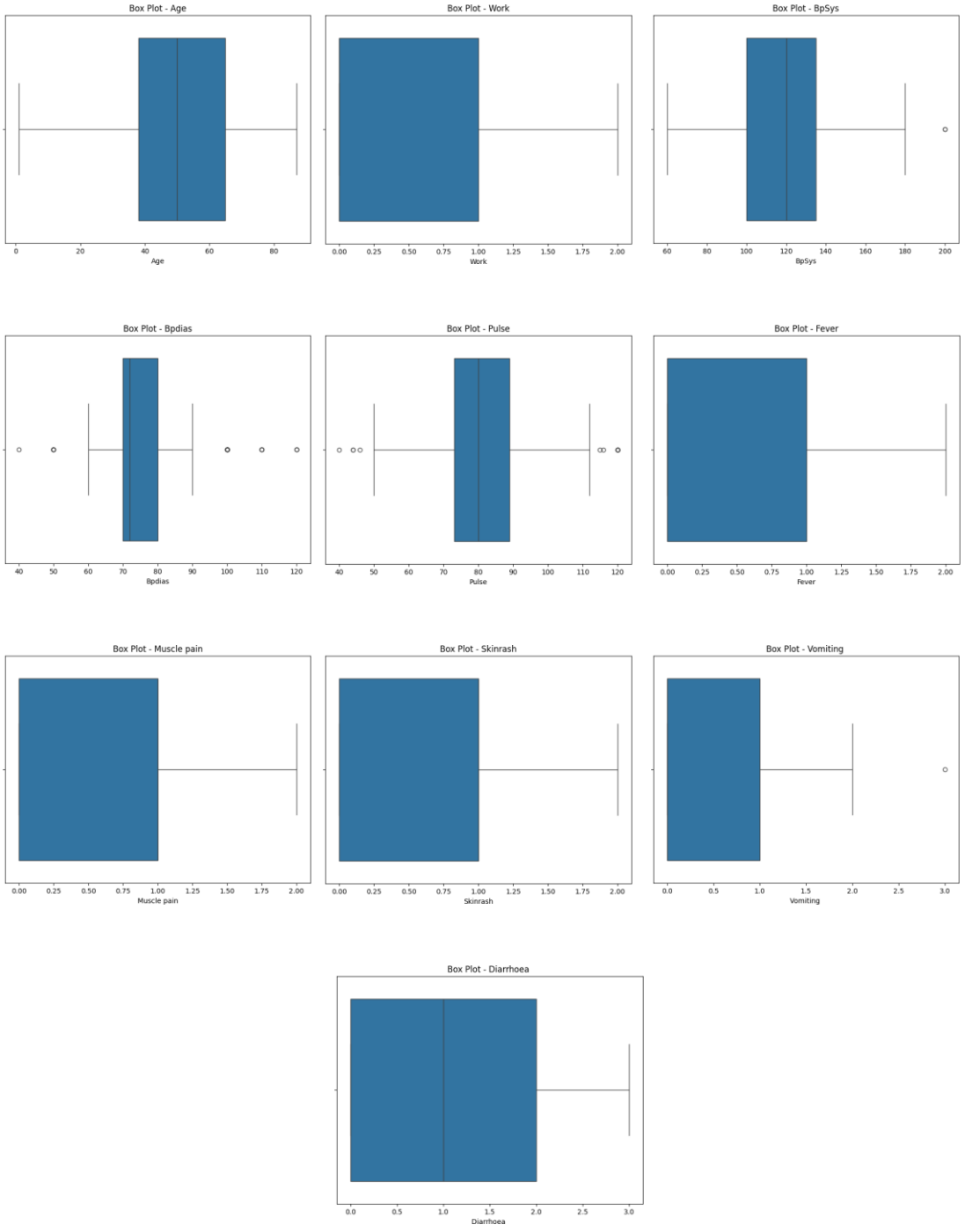


Fig. 4.10. Box plot for the Categorical Non Binary values

Below is the Histogram for all Parameters:

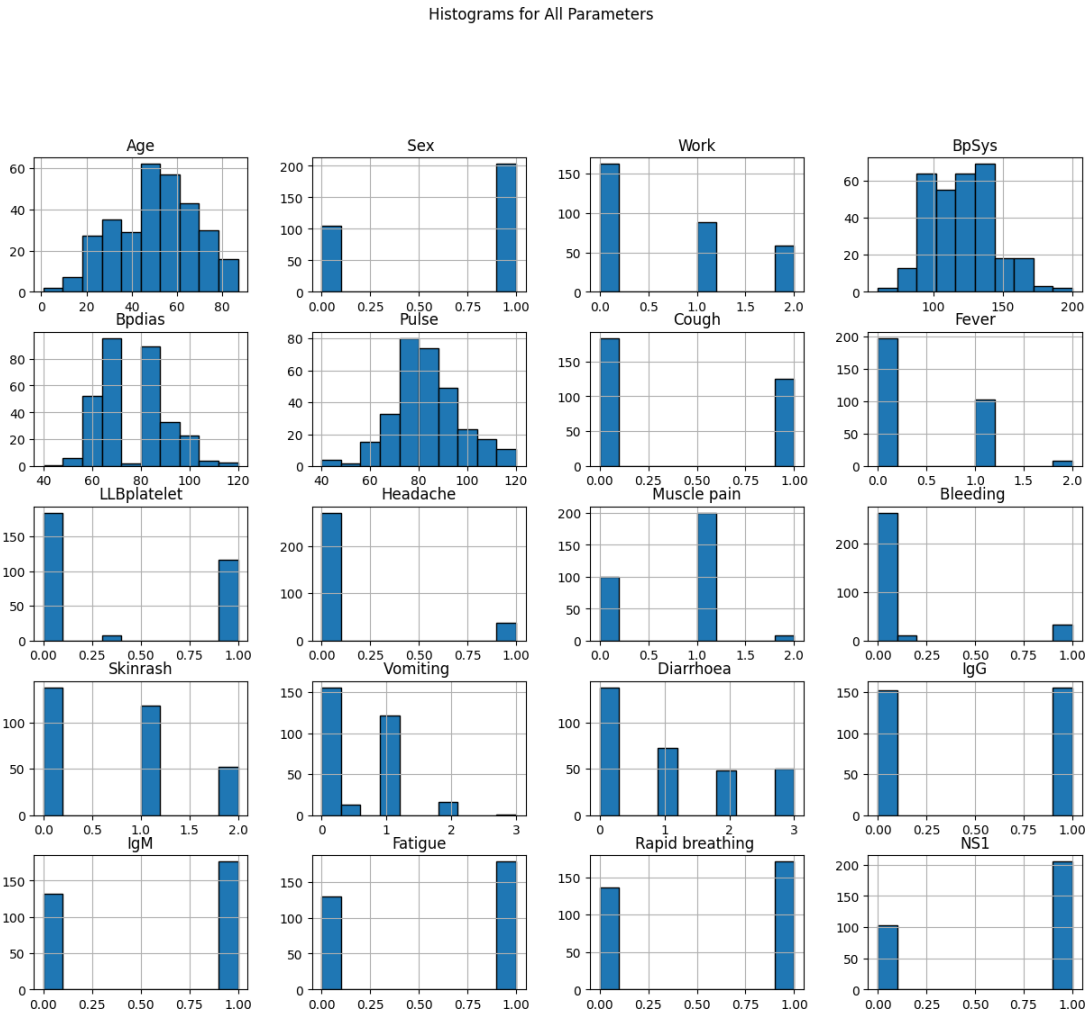


Fig. 4.11. Histogram for all Parameters

4.6 Model Evaluation

An Evaluation metrics for different models, a Confusion Matrix for the best model, an ROC Curve, and an AUC Curve are visualized below:

4.6.1 Comparison table for test accuracy of different classifiers:

Model	Test Accuracy
SVC	0.9677
Decision Tree	0.9677
XGBClassifier	0.9892
GaussianNB()	0.9892
KNeighbors	0.8280
Logistic Regression	0.9570
Linear Discriminant Analysis	0.9785
Random Forest	**0.9892**

TABLE 4.2: Comparison of Test Accuracies for Different Models

4.6.2 Evaluation metrics for different models:

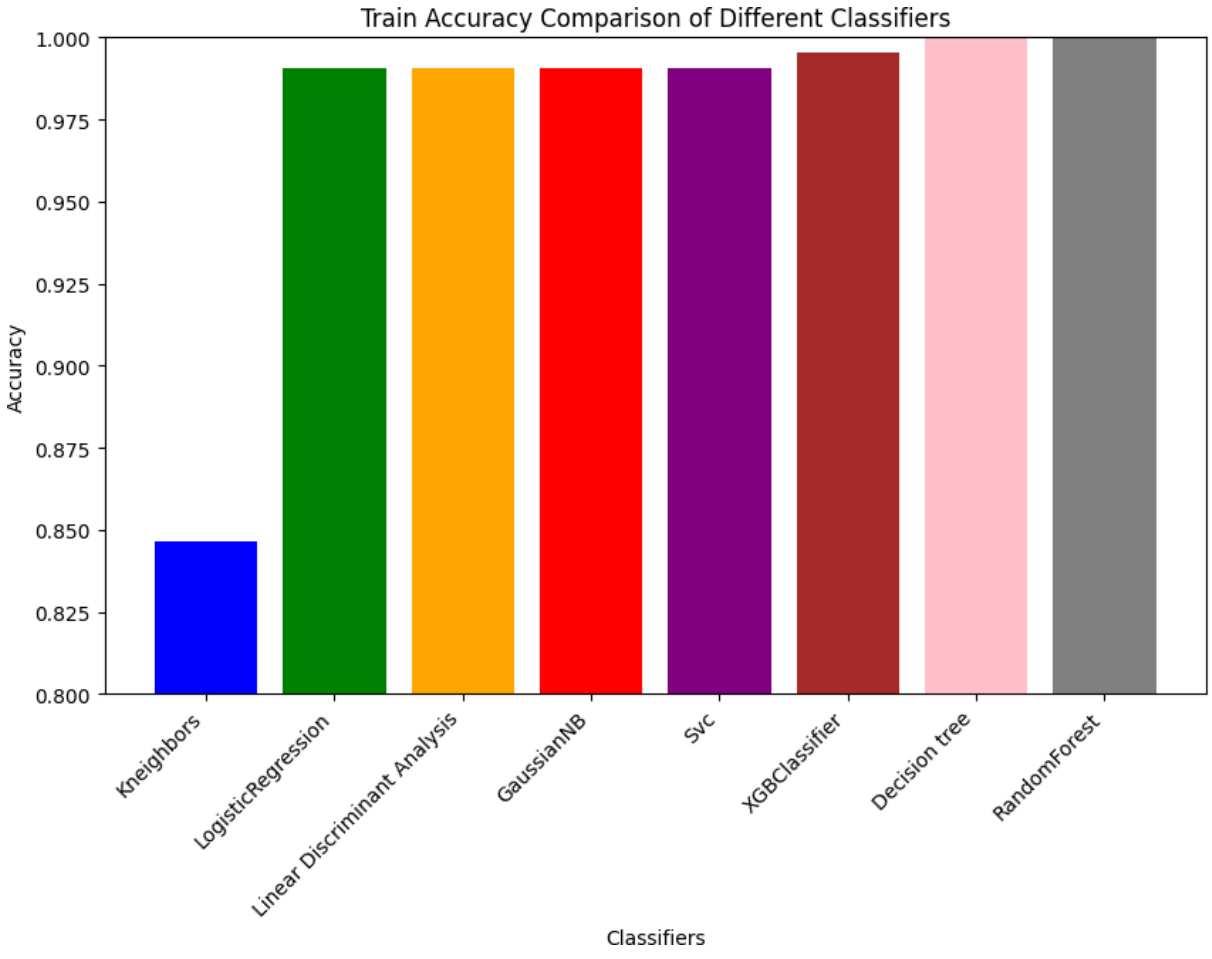


Fig. 4.12. Training Accuracy Comparison of different classifiers

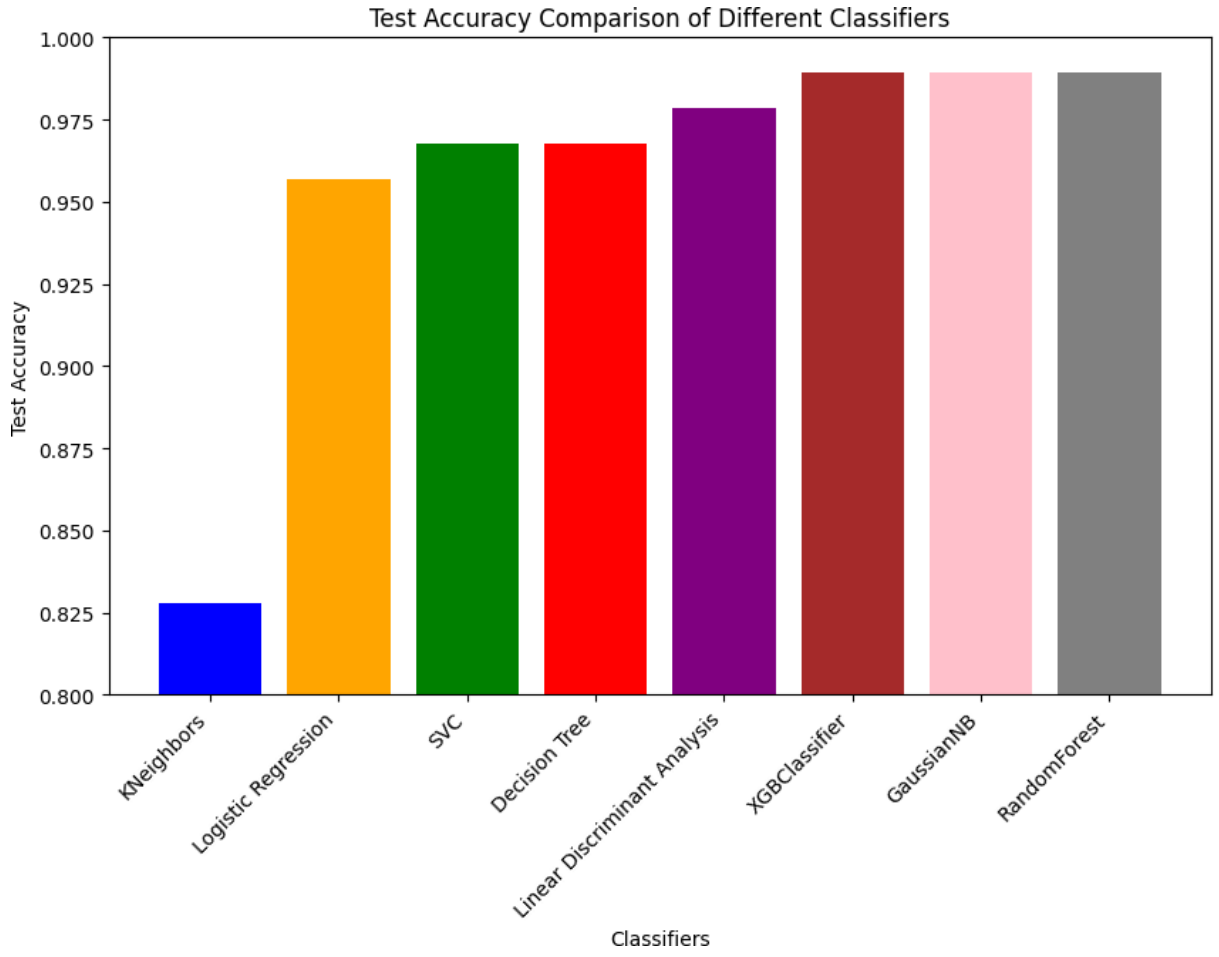


Fig. 4.13. Test Accuracy Comparison of different classifiers

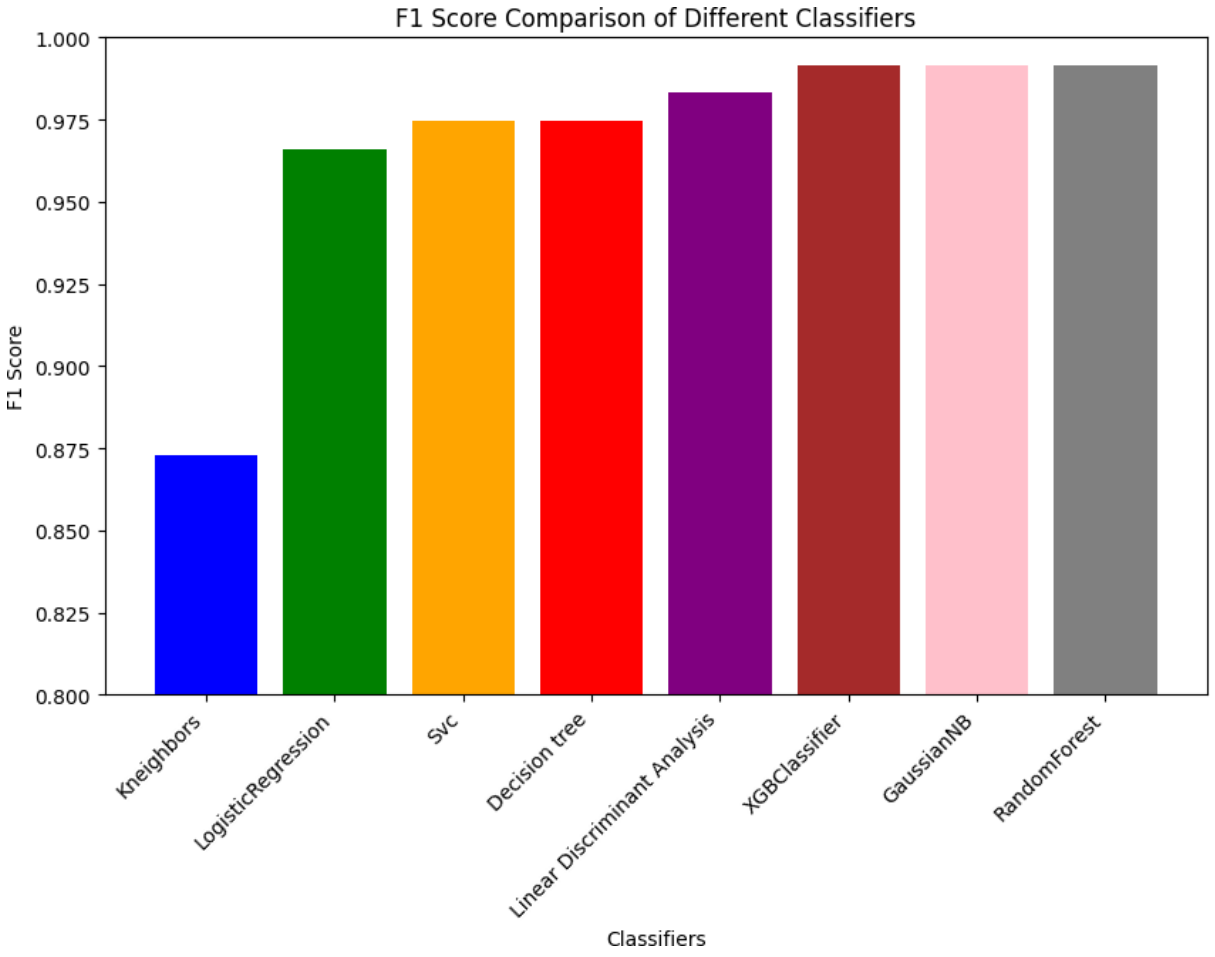


Fig. 4.14. F1 score comparison of different classifiers

4.6.3 Confusion Matrix for Random Forest:

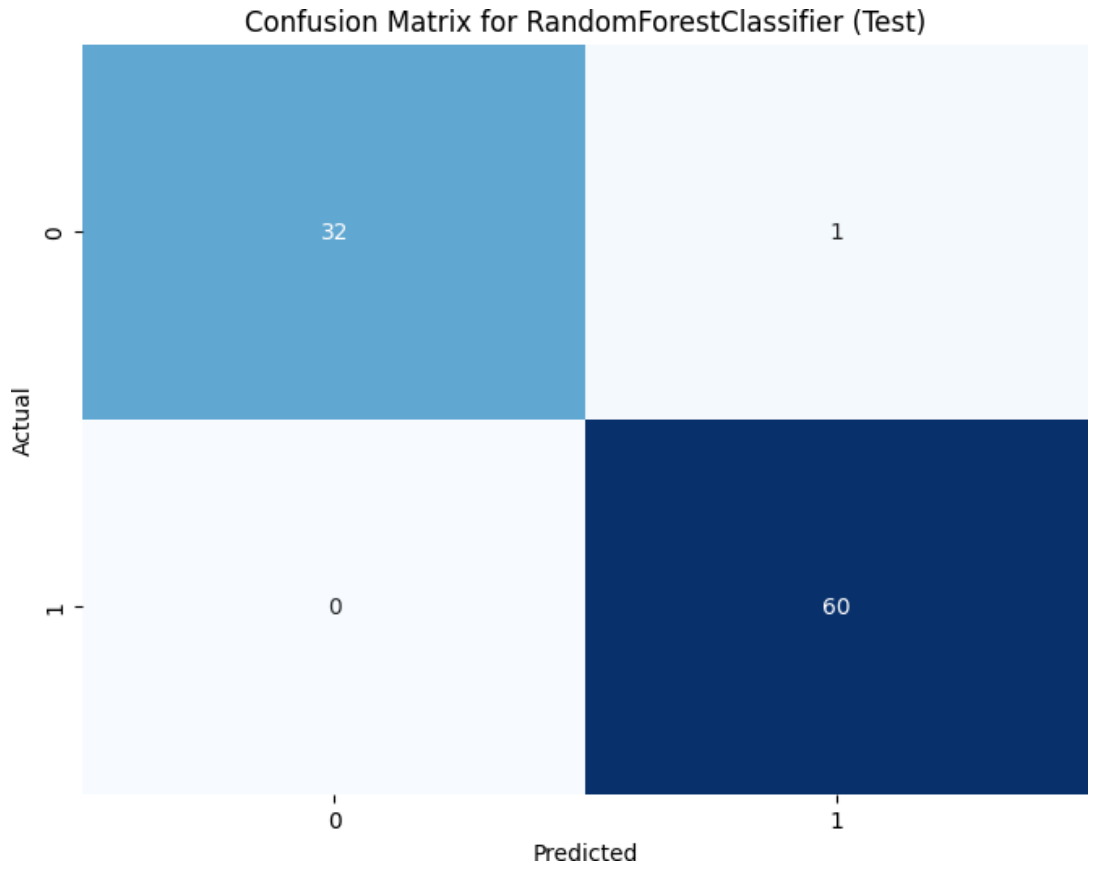


Fig. 4.15. Confusion Matrix for Random Forest

4.6.4 ROC Curve:

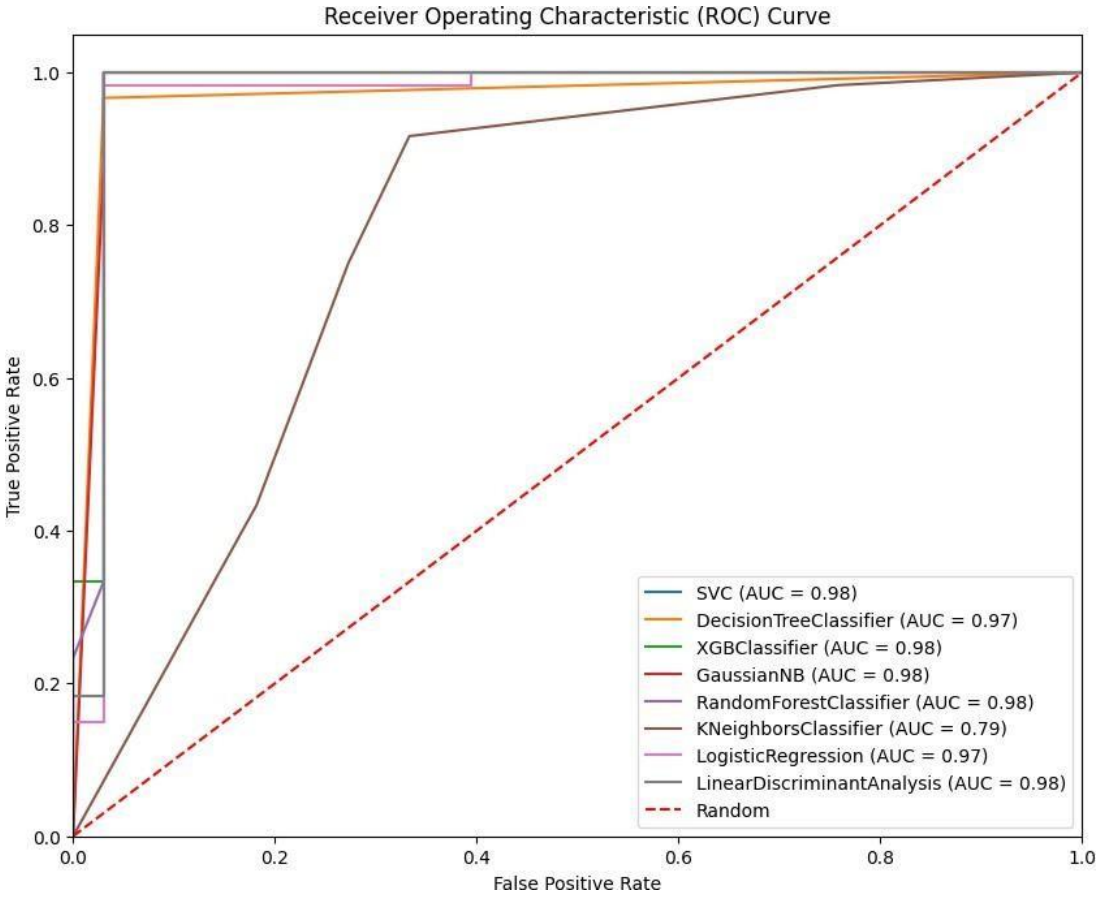


Fig. 4.16. ROC Curve

4.6.5 AUC Curve:

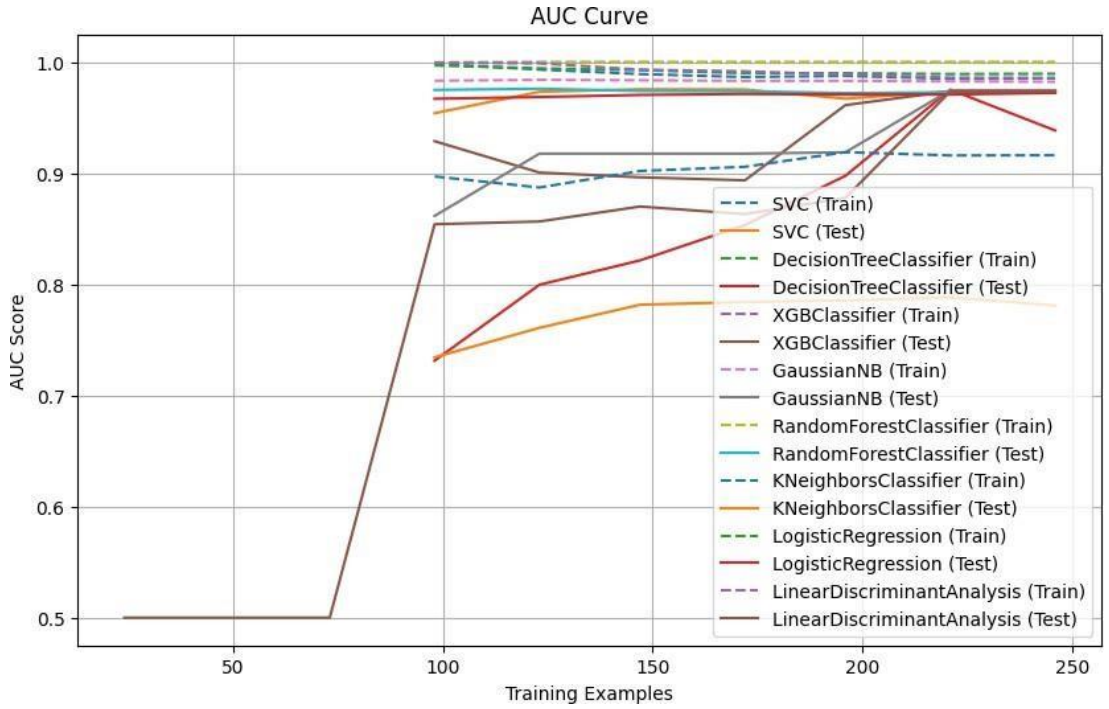


Fig. 4.17. AUC Curve

4.7 Demonstrating our model through app

We have demonstrated our model by building a web-based application using Streamlit

4.7.1 Screenshots of our web app:

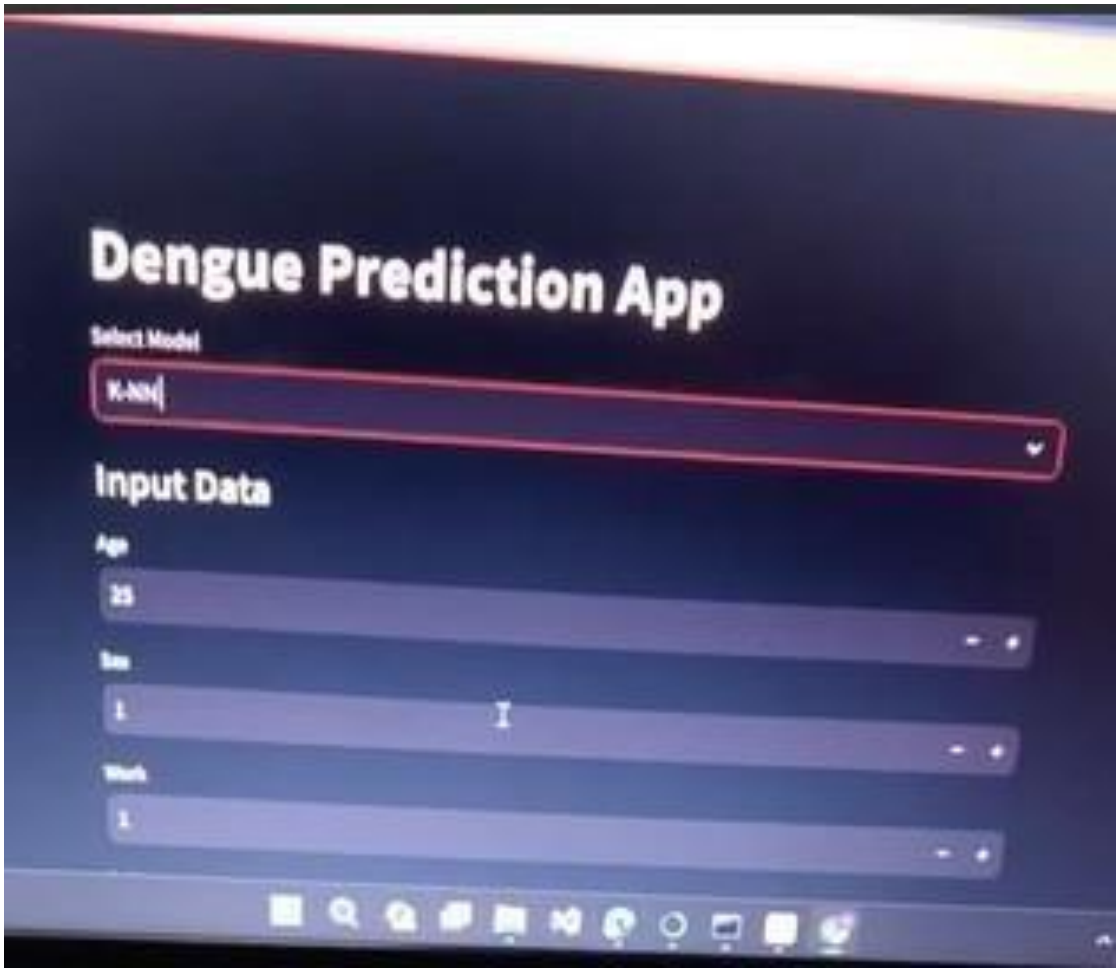


Fig. 4.18. Screenshot 1

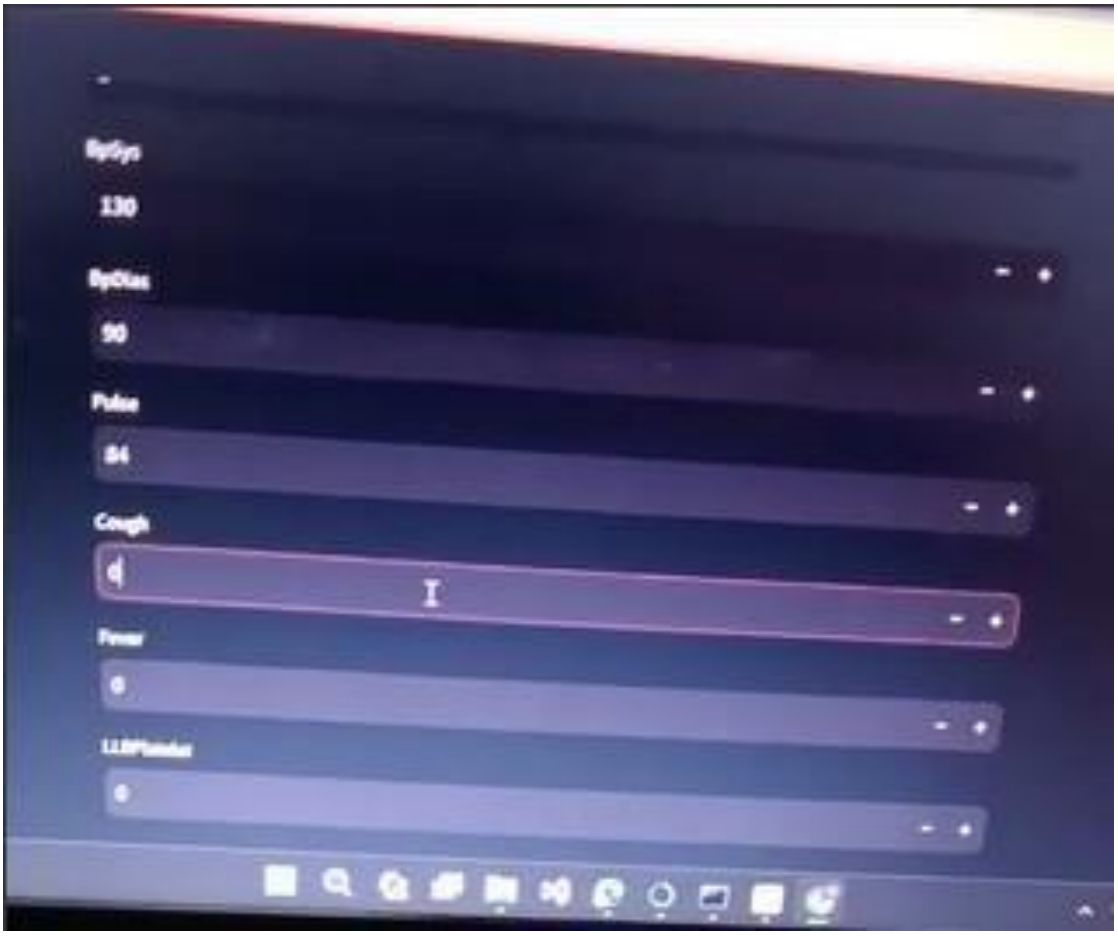


Fig. 4.19. Screenshot 2

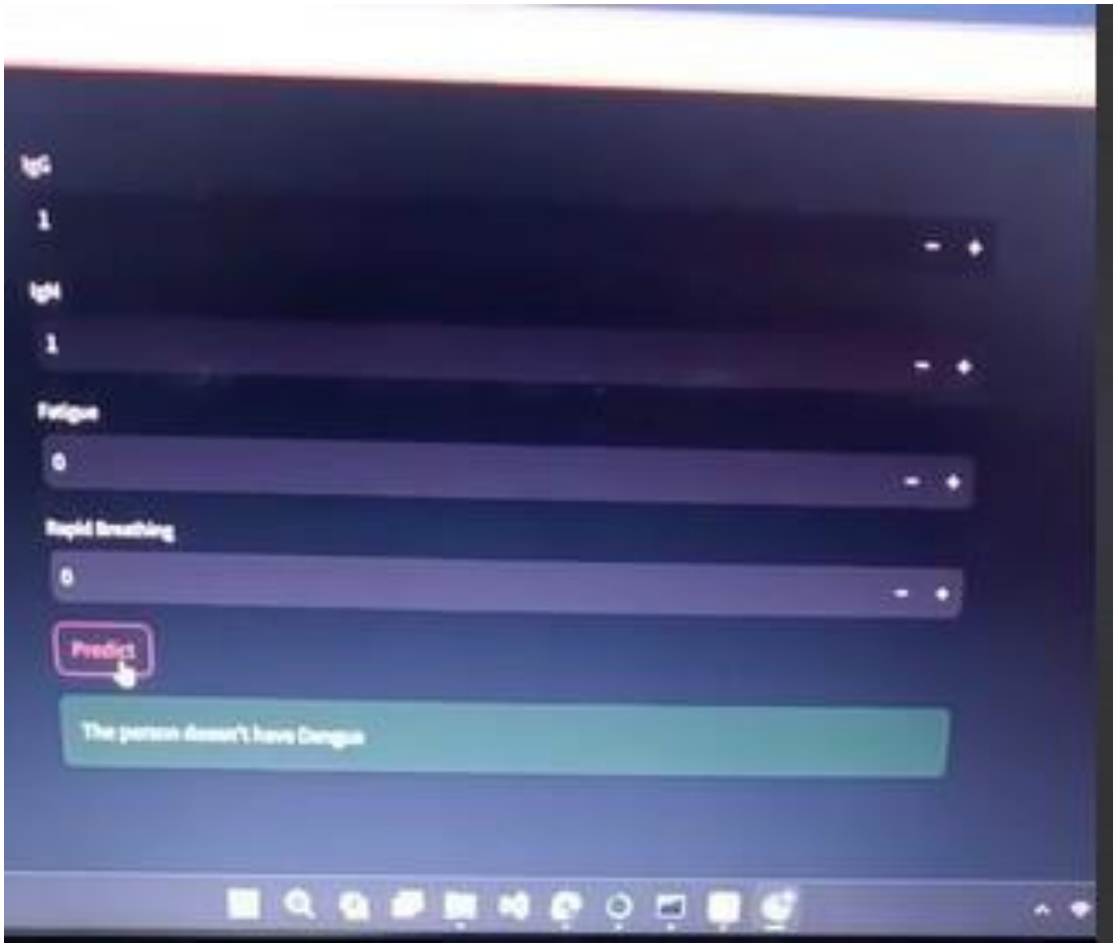


Fig. 4.20. Screenshot 3

4.8 Summary of Results

In this section, we highlighted all of our experimental results. We ran 8 classification models and showed their performance. Highlighting the Distribution of results, Correlation heatmap of dengue detection attributes, Count plot for the categorical binary values, Box plot for the categorical nonbinary values, Histogram, comparison table for test accuracy of different classifiers, evaluation metrics for different models, confusion matrix for random forest, ROC curve, AUC curve and lastly demonstrated in web app. After analysis, we can say that our model works smoothly and it will help to detect dengue detection cases accurately.

Chapter 5

Conclusion

We can conclude that our work has been completed with an accuracy of 98.92 percent. We have tested some other algorithms. The accuracy varies for iteration and is not better than the Random Forest. So, we selected the Random Forest model. Finally, we tried to detect dengue cases based on attributes.

5.1 Research Summary

We can see that few papers discussed local data and local diseases. In our work, we tried to focus on the field of dengue detection only. The discussed research papers authors have mentioned that they have used publicly available data from databases, or secondary data. However, we are using real-time data that are not available directly from databases or records from the healthcare industry.

5.2 Contribution of the work

- We have collected raw patients' data that would play a significant role in early dengue detection using machine learning techniques.
- We have performed ns1-based antigen testing using machine learning algorithms which would not only help in early and accurate detection of dengue but also provide better healthcare decision-making and finally bridge the gap between traditional clinical practices and modern computational approaches.
- We have built a web-based application that would provide a user-friendly interface to patients and healthcare professionals without any advanced technical knowledge

to test when immediate on-site testing is unavailable, would serve as an educational platform by providing related information to the disease, and would contribute significantly to timely decision-making in the diagnosis and management of dengue cases.

5.3 Future Work

- We will use more techniques like ns1-based antigen testing in association with machine learning in the future for early and accurate dengue detection.
- We will collect more patients' data and also add more attributes from different authentic sources to build a more detailed and predictive model.
- We will add geographical and climate-based data which will help in more precise dengue detection using machine learning algorithms.
- We will add a feedback option to our web-based app so that individuals interested can assist us in making the platform better through their suggestions. We will also add an open-source facility to our web-based app that would play an insignificant role in the detection of dengue by encouraging collaborative development from individuals interested, enabling the global community to enhance algorithms and features and thus contributing to more early and accurate detection of dengue using machine learning algorithms.

References

[1] R. Kapoor, V. Kadyan, and S. Ahuja, "Identification of influential parameter for early detection of dengue using machine learning approach," in *Proceedings of the 5th International Conference on Cyber Security & Privacy in Communication Net- works (ICCS)*, 2019.

[2] S. Amin, M. I. Uddin, D. H. AlSaeed, A. Khan, and M. Adnan, "Early detec- tion of seasonal outbreaks from twitter data using machine learning approaches," *Complexity*, vol. 2021, pp. 1–12, 2021.

[3] V. Janani, N. Maadhuryaa, D. Pavithra, and S. R. Sree, "Dengue prediction using (mlp) multilayer perceptron—a machine learning approach," *Int. J. Res. Eng. Sci. Manag*, vol. 3, pp. 226–231, 2020.

[4] N. A. M. Salim, Y. B. Wah, C. Reeves, M. Smith, W. F. W. Yaacob, R. N. Mudin, R. Dapari, N. N. F. F. Sapri, and U. Haque, "Prediction of dengue outbreak in selangor malaysia using machine learning techniques," *Scientific reports*, vol. 11, no. 1, p. 939, 2021.

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

[6] S. Q. Ong, P. Isawasan, A. M. M. Ngesom, H. Shahar, A. m. M. Lasim, and G. Nair, "Predicting dengue transmission rates by comparing different machine learning models with vector indices and meteorological data," *Scientific reports*, vol. 13, no. 1, p. 19129, 2023.

[7] T. Sajana, M. Navya, Y. Gayathri, and N. Reshma, "Classification of dengue using machine learning techniques," *Int J Eng Technol*, vol. 7, no. 2.32, pp. 212–218, 2018.

[8] J. K. Chaw, S. H. Chaw, C. H. Quah, S. Sahrani, M. C. Ang, Y. Zhao, and T. T. Ting, "A predictive analytics model using machine learning algorithms to estimate

the risk of shock development among dengue patients," *Healthcare Analytics*, vol. 5, p. 100290, 2024.

[9] S. Chattopadhyay, A. Chattopadhyay, and E. Aifantis, "Predicting case fatality of dengue epidemic: Statistical machine learning towards a virtual doctor," *Journal of Nanotechnology in Diagnosis and Treatment*, vol. 7, pp. 10–24, 2021.

[10] Y. E. Liu, S. Saul, A. M. Rao, M. L. Robinson, O. L. Agudelo Rojas, A. M. Sanz, M. Verghese, D. Solis, M. Sibai, C. H. Huang, *et al.*, "An 8-gene machine learning model improves clinical prediction of severe dengue progression," *Genome medicine*, vol. 14, no. 1, p. 33, 2022.

[11] D. Salami, C. A. Sousa, M. d. R. O. Martins, and C. Capinha, "Predicting dengue importation into europe, using machine learning and model-agnostic methods," *Scientific Reports*, vol. 10, no. 1, p. 9689, 2020.

[12] P. Silitonga, B. E. Dewi, A. Bustamam, and H. S. Al-Ash, "Evaluation of dengue model performances developed using artificial neural network and random forest classifiers," *Procedia Computer Science*, vol. 179, pp. 135–143, 2021.

[13] S.-W. Huang, H.-P. Tsai, S.-J. Hung, W.-C. Ko, and J.-R. Wang, "Assessing the risk of dengue severity using demographic information and laboratory test results with machine learning," *PLoS neglected tropical diseases*, vol. 14, no. 12, p. e0008960, 2020.

[14] S. G. Kakarla, P. K. Kondeti, H. P. Vavilala, G. S. B. Boddeda, R. Mopuri, S. Kumaraswamy, M. R. Kadiri, and S. R. Mutheneni, "Weather integrated multiple machine learning models for prediction of dengue prevalence in india," *International Journal of Biometeorology*, vol. 67, no. 2, pp. 285–297, 2023.

[15] E. Mussumeci and F. C. Coelho, "Machine-learning forecasting for dengueepidemics-comparing lstm, random forest and lasso regression," *Medrxiv*, pp. 2020–01, 2020.

[16] A. Appice, Y. R. Gel, I. Iliev, V. Lyubchich, and D. Malerba, "A multi-stage machine learning approach to predict dengue incidence: a case study in mexico," *Ieee Access*, vol. 8, pp. 52713–52725, 2020.

[17] S. Saturi, "Development of prediction and forecasting model for dengue disease using machine learning algorithms," in *2020 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*, pp. 6–11, IEEE, 2020.

[18] F. Yavari Nejad and K. D. Varathan, "Identification of significant climatic risk factors and machine learning models in dengue outbreak prediction," *BMC medical informatics and decision making*, vol. 21, no. 1, pp. 1–12, 2021.

[19] A. N. A. Kamarudin, Z. Zainol, and N. F. A. Kassim, "Forecasting the dengue outbreak using machine learning algorithm: A review," in *2021 International Conference of Women in Data Science at Taif University (WiDSTaif)*, pp. 1–5, IEEE, 2021.

[20] A. Aziz and A. Aziz, "Dengue cases prediction using machine learning approach," *iRASD Journal of Computer Science and Information Technology*, vol. 2, no. 1, pp. 13–25, 2021.

[21] L. Tuan, M.-T. Vo, T. Pham, and S. Dao, "A novel wrapper–based feature selection for early diabetes prediction enhanced with a metaheuristic," *IEEE Access*, vol. 9, pp. 7869 – 7884, 12 2020.