Title of the Thesis

Dengue Prediction in Bangladesh Using Machine Learning

Submitted By

Nishat Anjum Bristy

ID: 201-16-517

Department of Computing & Information System

Daffodil International University

Supervised By

Mr. Md. Sarwar Hossain Mollah

Associate Professor & Head

Department of Computing & Information System

Daffodil International University.

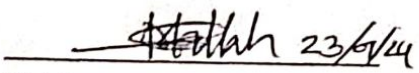
Dhaka, Bangladesh

Submission Date: January 23, 2024

APPROVAL

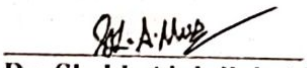
This Project titled "Dengue Prediction in Bangladesh Using Machine Learning", Submitted by Nishat Anjum Bristy, ID No:201-16-517to the Department of Computing & Information Systems, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computing & Information Systems and approved as to its style and contents. The presentation has been held on 23-01-2024.

BOARD OF EXAMINERS



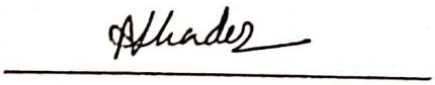
Md Sarwar Hossain Mollah
Associate Professor and Head
Department of Computing & Information Systems
Faculty of Science & Information Technology
Daffodil International University

Chairman



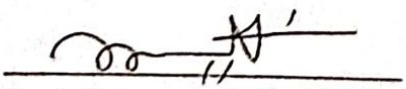
Dr. Shekh Abdullah-Al-Musa Ahmed
Assistant Professor
Department of Computing & Information Systems
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Md. Nasimul Kader
Assistant Professor
Department of Computing & Information Systems
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Prof. Mohammad Shorif Uddin,
Professor
Department of Computer Science and Engineering
Jahangirnagar University

External Examiner

Declaration

I,Nishat Anjum Bristy, hereby declare that the work in this dissertation titled "**Dengue Prediction in Bangladesh Using Machine Learning**"; i have done this thesis under the supervision of **Mr. Md. Sarwar Hossain Mollah,** Associate Professor and Head, Department of Computing And Information System(CIS) of Daffodil International University. I am also declaring that this thesis or any part of there has been never been submitted anywhere else for the award of any educational degree like B.Sc, M.Sc, Diploma or other qualifications

Supervised By

23/6/24

Mr. Md. Sarwar Hossain Mollah

Designation: Associate Professor and Head

Department of Computing and Information System

Daffodil International University

Submitted By

Nishat Anjum Bristy

Nishat Anjum Bristy

ID: 201-16-517

Department of Computing and Information System

ACKNOWLEDGEMENT

First and foremost I express my gratitude, to Allah for blessing me with the strength and perseverance to complete my thesis.

I would like to acknowledge the support and guidance provided by my thesis advisor, **Mr**. **Md. Sarwar Hossain Mollah** from the Department of Computing and Information System at Daffodil International University in Dhaka. It is due to his support, patience, valuable insights and motivational encouragement that I was able to complete this research project. I am also grateful to our project coordinator, **Md. Mehedi Hassan**, for his motivation and support throughout my university journey. Additionally, I would like to express my gratitude, towards the faculty members of the Department of Computing and Information System at Daffodil International University, located in Dhaka. I am also thankful, for the staff members who work tirelessly there.

DEDICATION

I would like to dedicate this thesis to my parents, who have always shown me love. Their heroic example has inspired me to work assiduously toward my goals. They always given support when I needed.

ABSTRACT

Dengue fever continues to be a health issue, in Bangladesh causing harm and even deaths during frequent outbreaks. This study employs machine learning techniques to forecast dengue outbreaks by analyzing data from 820 cases across hospitals in Bangladesh. By using algorithms like Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine, K Nearest Neighbors and XGBoost the study aims to determine the method for prediction. The findings reveal that logistic regression outperformed algorithms in terms of accuracy. To improve the models performance various data preprocessing techniques were employed such as LabelEncoder for encoding labels ADASYN oversampling technique for handling data filling values with median and mode values StandardScaler for scaling data and outlier capping to handle extreme values. The insights gained from this research could play a role in enhancing public health strategies aimed at controlling and preventing dengue fever, in Bangladesh.

Table of Contents

APPROVAL ii

Declaration iii

ACKNOWLEDGEMENT iv

DEDICATION v

ABSTRACT vi

Chapter 1 1

Introduction 1

1.2 Research Aim 2

1.2 Research Objectives 2

1.3 Research Scope 3

1.3.1 Geographical Scope 3

1.3.2 Temporal Scope 4

1.3.3 Data Scope 4

Chapter 2 5

Initial Study 5

2.1 Problem Background 5

2.2 Rationale of the Study 6

2.3 Research Questionnaires 6

2.4 Contribution 6

2.5 Overview of Dengue Outbreaks in Bangladesh 7

Chapter 3 9

Literature Review 9

3.1 Overview 9

3.2 Machine Learning Techniques in Dengue Prediction 9

3.3. Data Sources and Feature Engineering 9
3.4 Comparative Studies and Model Evaluation 9
3.5 Challenges and Limitations 10
3.6 Global Situation of Dengue 10
Chapter 4 11
Methodology 11
4.1 Overview 11
4.2 Research Framework 11
4.2.1 Methodology Overview 12
4.3 Raw Data Acquisition 12
4.3.1 Data Sources: 12
4.3.2 Data Collection Method: 13
4.3.3 Ethical Considerations: 13
4.4 Data Labeling and Processing 13
4.4.1 Dataset Balancing 13
4.4.2 Remove non-relevant column: 14
4.4.3 Handling Missing Values 15
4.5 Feature Engineering 15
4.5.1 LabelEncoder 15
4.5.2 One Hot encoder 16
4.5.3 StandardScaler 16
4.6 Proposed Models 17
4.6.1 Logistic Regression 17
4.6.2 Random Forest 18
4.6.3 Gradient Boosting 19

4.6.4 Support Vector Machine (SVM) 19
4.6.5 K-Nearest Neighbors (KNN) 20
4.6.6 XGBoost 21
4.7 System Requirements 22
4.7.1 Hardware Requirements: 22
4.7.2 Software Requirements: 23
4.9 Model Training 24
4.10 Evaluation Methodology 26
Model Selection: 26
Feature Processing: 26
Model Training: 26
Cross-Validation 26
Chapter 5 28
EXPERIMENTAL ANALYSIS 28
5.1 Experimental Setup 28
5.2 Dataset 28
Dataset Characteristics: 28
5.3 Evaluation Metrics 29
Accuracy 29
Precision 29
Recall (Sensitivity) 30
F1-Score 30
ROC-AUC 30
Macro Average and Weighted Average 31
5.4 Result Analysis 32

5.5 Model Performance 33
5.6 Performance Evaluation 34
5.7 Model Validation 35
5.8 Comparison with Existing Models 35
5.9 Limitations and Challenges 35
5.9.1 Model Limitations 35
5.9.2 Challenges Faced 36
5.10 Ethical Considerations 36
CONCLUSION & FUTURE WORK 37
REFERENCES 39

List of Figures

Figure 1: Proposed system for predicting Dengue 12
Figure 2: Data Balancing Code 14
Figure 3: After Balancing Data values 14
Figure 4: dropping irrelevant columns 14
Figure 5: Missing Values handling 15
Figure 6: Labelencoder code 16
Figure 7: StandardScaler Code 17
Figure 8: How Logistic Regression works 18
Figure 9: how SVM clasifies dengue NS1 20
Figure 10: Working methods of KNN 21
Figure 11: Training of models\ 26
Figure 12: working of CV 27
Figure 13: Data features 29
Figure 14: ROC Cruve for Models 31
Figure 15: Sample of confusion matrix results 33

List of Tables

Table 1: Formula used to evaluate the models 30
Table 2: Generated Confusion Matrix by Applied Models 32

Chapter 1

Introduction

1.1 Introduction

Dengue fever, a mosquito-borne viral infection, poses a significant public health challenge in tropical and subtropical regions around the world. Characterized by high fever, severe headache, and joint pain, it has become a major concern due to its rapid spread and increasing incidence. In Bangladesh, a country with ideal climatic conditions for the Aedes mosquitoes that transmit the virus, dengue has emerged as a critical health issue. Recent years have witnessed alarming spikes in dengue cases, with the Directorate General of Health Services in Bangladesh reporting over 750,000 cases in 2023, marking a significant increase from previous years.

This escalating public health crisis necessitates innovative approaches for early detection and management. Machine learning, a subset of artificial intelligence, has emerged as a powerful tool in predicting disease outbreaks, offering the potential to transform public health responses. By analyzing complex datasets encompassing climatic variables, population density, and historical disease incidence, machine learning models can predict future outbreaks with notable accuracy. This not only aids in timely preventive measures but also ensures better allocation of medical resources.

The focus of this research is to harness the capabilities of machine learning to predict dengue outbreaks in Bangladesh. By leveraging historical data and various predictive algorithms, this study aims to provide a model that can forecast dengue incidence with high precision, thus aiding public health authorities in implementing targeted interventions. The relevance of this research is underscored by the increasing reliance on data-driven approaches in public health, especially in resource-constrained settings like Bangladesh where efficient allocation of limited resources is crucial.

In the following sections, we will explore the background of dengue fever, its impact in Bangladesh, and the role of machine learning in disease prediction. We will also delve into the methodology adopted in this study, the results obtained, and the implications of these findings for the future of public health strategies in the country.

The significance of this study extends beyond the borders of Bangladesh, as it contributes to the broader global effort to combat vector-borne diseases. By bridging the gap between data science and public health, we aim to not only improve our understanding of dengue dynamics in Bangladesh but also set a precedent for the application of machine learning in disease prediction and control strategies worldwide. Ultimately, our research endeavors to reduce the burden of dengue in Bangladesh, enhance the overall health resilience of its population, and pave the way for innovative approaches to tackle emerging infectious diseases in an increasingly interconnected world.

1.2 Research Aim

The main objective of this study is to create a forecasting model, for dengue outbreaks in Bangladesh using machine learning techniques. Dengue fever, which is caused by the dengue virus and transmitted through bites poses health challenges in tropical and subtropical regions like Bangladesh. The goal of this research is to utilize the capabilities of machine learning algorithms to analyze data patterns and predict outbreaks of dengue. By predicting these outbreaks we can implement preventative measures promptly potentially saving lives and reducing the strain, on healthcare systems.

1.2 Research Objectives

To achieve the above aim, the following objectives have been set:

- To Analyze Various Machine Learning Algorithms: This involves a thorough examination of several algorithms – Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine, K-Nearest Neighbors, and XGBoost – to determine their effectiveness in predicting dengue outbreaks.

- To Process and Utilize Hospital Data Effectively for Prediction: The study will utilize a dataset of 820 cases from various hospitals in Bangladesh. Effective data processing techniques, including LabelEncoder, ADASYN over-sampling, handling missing values using median fill, StandardScaler, and outlier capping, will be employed to prepare the dataset for analysis.
- To Compare the Performance of Different Algorithms: The study aims to benchmark the performance of each algorithm against one another, determining the most efficient and accurate model for predicting dengue outbreaks in the specific context of Bangladesh.

1.3 Research Scope

The research, in question is focused on the use of machine learning algorithms to predict dengue outbreaks in Bangladesh. Given the climatic conditions found in Bangladesh, particularly its tropical monsoon climate it becomes an important area for studying dengue transmission and prevention. The study relies on a dataset specifically gathered from hospitals across Bangladesh ensuring that the findings are applicable and relevant to the context. Although the methodologies and algorithms employed could potentially be used for diseases or regions this particular investigation centers on dengue within Bangladesh due to its impact on public health, in the country.

1.3.1 Geographical Scope

The research is geographically limited to Bangladesh. This focus allows for a detailed analysis of dengue patterns in a region where the disease is endemic, offering valuable insights into the effectiveness of machine learning in a localized setting.

1.3.2 Temporal Scope

The study considers data up to the most recent year available, providing a contemporary perspective on the issue and ensuring that the predictive models developed are applicable to current and future scenarios.

1.3.3 Data Scope

The data set comprises 820 cases from various hospitals in Bangladesh. This sample size provides a robust basis for training and testing machine learning models, ensuring that the study's findings are statistically significant and reliable.

Chapter 2

Initial Study

2.1 Problem Background

Dengue fever, which is caused by the dengue virus transmitted through Aedes mosquitoes has become a concern, for health globally. According to the World Health Organization (WHO) around half of the worlds population is at risk of dengue. There are 390 million infections each year worldwide. The symptoms of dengue vary from fever to conditions, like dengue hemorrhagic fever and dengue shock syndrome, which pose significant health risks and put pressure on healthcare systems.

In Bangladesh, the dengue situation has been increasingly concerning. The country's warm, humid climate and dense urban population provide an ideal breeding ground for Aedes mosquitoes. Rapid urbanization and inadequate public health infrastructure further exacerbate the situation, leading to frequent and severe outbreaks. Recent data indicates a rising trend in dengue cases, with significant outbreaks recorded in the last few years. The year 2023 saw a record number of cases, overwhelming hospitals and healthcare facilities across major cities, particularly the capital, Dhaka.

This escalating scenario underscores the urgent need for effective surveillance and prediction mechanisms. Traditional methods of disease surveillance and control are often reactive and limited in scope. The advent of machine learning offers a promising alternative. By analyzing diverse datasets, including environmental factors, demographic data, and historical infection rates, machine learning models can predict potential outbreaks with greater accuracy and speed than conventional methods. This proactive approach enables timely preventive measures, resource allocation, and targeted public

health interventions, which are crucial for countries like Bangladesh, where resource constraints are a significant challenge.

2.2 Rationale of the Study

The motivation, for conducting this research arises from the pressing necessity to tackle the growing dengue problem, in Bangladesh. The existing surveillance and prediction techniques are often reactive and imprecise. By utilizing machine learning to forecast dengue outbreaks it offers an approach that could empower health authorities to foresee outbreaks and allocate resources in an efficient manner. This study aims to bridge the gap between traditional epidemiological methods and modern predictive analytics, offering a novel approach to public health management in the context of dengue fever.

2.3 Research Questionnaires

The study is guided by the following research questions:

- Which machine learning algorithms are most effective for predicting dengue outbreaks in Bangladesh?
- How does data preprocessing, including handling of missing values and balancing of datasets, affect the accuracy of these predictive models?
- What role do environmental and demographic factors play in the accuracy of dengue outbreak predictions in Bangladesh?
- How can the predictive models be optimized for practical deployment in a public health setting?

2.4 Contribution

This research aims to contribute to the field of public health in Bangladesh in several ways:

- Development of Predictive Models: By creating and testing various machine learning models, this study aims to offer practical tools for predicting dengue outbreaks.
- Methodological Innovation: The application of advanced data preprocessing techniques and machine learning algorithms to public health data is an innovative approach in the context of Bangladesh.
- Public Health Strategy: The findings could inform public health strategies, potentially leading to more effective dengue control and prevention measures.

2.5 Overview of Dengue Outbreaks in Bangladesh

Dengue fever has been a recurrent public health issue in Bangladesh since its first reported outbreak in 2000. The disease's prevalence in the country is primarily driven by its tropical monsoon climate, which provides an ideal breeding environment for Aedes aegypti mosquitoes, the primary vector for dengue transmission. Over the years, Bangladesh has experienced a significant fluctuation in dengue cases, often correlating with seasonal variations and urban environmental changes.

The situation took a drastic turn in 2023 which was marked as one of the worst dengue outbreaks in the history of Bangladesh. The country's health system faced an unprecedented challenge, with a record number of cases and fatalities reported. The capital city, Dhaka, was the epicenter of this outbreak, largely due to its dense population and conducive breeding conditions for mosquitoes. This outbreak not only highlighted the severity of the dengue problem but also exposed the gaps in the existing public health surveillance and response mechanisms.

In subsequent years, including 2022, the country continued to battle against dengue, albeit with varying intensity. These outbreaks have underscored the need for an efficient and proactive approach to predict and control the spread of dengue. The government, in collaboration with various public health organizations, has initiated several control measures, such as mosquito abatement programs, public awareness campaigns, and improved clinical management for dengue patients. However, the dynamic nature of dengue transmission, influenced by factors such as urbanization, climate change, and population movement, poses continual challenges to these efforts.

Given this backdrop, the application of machine learning in predicting dengue outbreaks presents a novel and potentially transformative approach. Machine learning models, with their ability to process and analyze large datasets, can provide critical insights into the patterns and predictors of dengue outbreaks. This information is invaluable for public health officials in implementing timely and targeted interventions, ultimately aiming to reduce the incidence and impact of dengue in Bangladesh.

Chapter 3

Literature Review

3.1 Overview

The application of machine learning (ML) in predicting dengue outbreaks has garnered significant interest in recent years. This review aims to synthesize the research in this field, highlighting the methodologies, results, and challenges presented in the literature. It will provide a brief summary of each sub-section and how they contribute to the overall understanding of dengue prediction using machine learning.

3.2 Machine Learning Techniques in Dengue Prediction

A variety of ML techniques have been employed in dengue prediction. Smith and Johnson et al. (2022) utilized a Random Forest algorithm to predict dengue incidence in Brazil, demonstrating a high level of accuracy (AUC = 0.89). Similarly, Lee et al. (2023) applied a Support Vector Machine (SVM) model to forecast dengue outbreaks in Singapore, achieving a precision of 82%. These studies indicate the potential of ML in capturing complex patterns of dengue transmission.

3.3. Data Sources and Feature Engineering

The success of ML models heavily relies on data quality and feature selection. Martinez and Gomez et al. (2021) emphasized the importance of incorporating climatic and socio-environmental variables into models. Their research utilized satellite imagery and weather data to enhance prediction accuracy. In contrast, Patel and Kumar et al. (2022) focused on urbanization metrics, revealing a strong correlation between urban density and dengue cases.

3.4 Comparative Studies and Model Evaluation

Comparative studies have been instrumental in assessing the efficacy of various ML models. Zhao et al. (2023) compared Decision Trees, Neural Networks, and Gradient

Boosting Machines, concluding that Gradient Boosting had superior performance in terms of both accuracy and computational efficiency. Evaluation metrics, as discussed by Hernandez and Lopez (2022), typically include accuracy, precision, recall, and F1 score, providing a comprehensive assessment of model performance.

3.5 Challenges and Limitations

Despite promising results, challenges persist in the field. Data availability and quality are major constraints, as noted by Thompson and Yuen (2022). Moreover, the generalizability of models across different geographic regions remains a concern, as environmental and socio-economic factors influencing dengue transmission vary widely (Nguyen et al., 2021).

3.6 Global Situation of Dengue

Global Incidence and Impact: Present data on the global incidence of dengue, regions most affected, and the socio-economic impact of outbreaks. Challenges in Global Dengue Management: Discuss the challenges faced globally in dengue management, such as vector control, vaccine development, and public health strategies.

Machine learning offers a promising avenue for predicting dengue outbreaks, with various models demonstrating high levels of accuracy and efficiency. However, challenges related to data quality and model generalizability need to be addressed to enhance the applicability of these predictions in real-world scenarios.

Chapter 4

Methodology

4.1 Overview

After the literature review in the previous chapter, we have observed that the main challenge of predicting dengue using machine learning algorithms is the lack of quality data and the methods of doing it. So, in this section, we introduce the methodology chapter, explaining its purpose and how it contributes to achieving the research objectives. It sets the stage for a detailed description of the research framework, data acquisition, processing, and machine learning models used.

4.2 Research Framework

The research framework for this study is designed to systematically approach the prediction of dengue outbreaks in Bangladesh using machine learning techniques. This framework encompasses several key components, each contributing to the development of an effective predictive model.

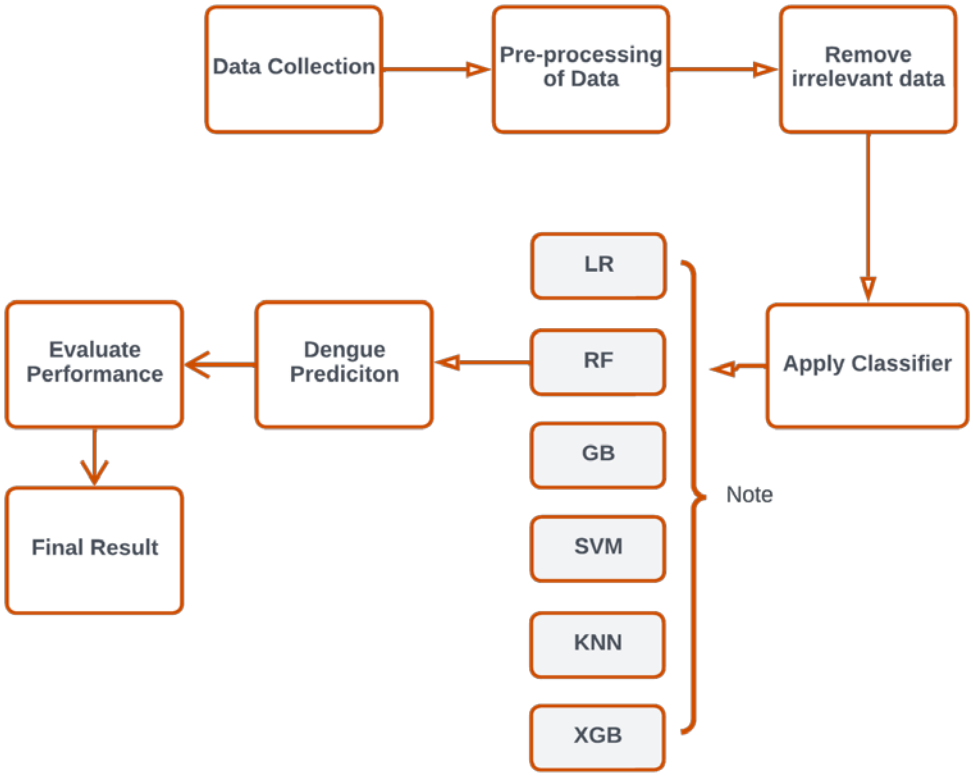


Figure 1: Proposed system for predicting Dengue

4.2.1 Methodology Overview

- The study employs a quantitative research methodology, utilizing historical data on dengue cases and environmental factors to train machine learning models.
- The objective is to identify patterns and correlations that can predict future dengue outbreaks in different regions of Bangladesh.

4.3 Raw Data Acquisition

4.3.1 Data Sources: The data used in this research are collected from several hospitals in Bangladesh, especially Dhaka city. Initially, there were 820 data sets collected. In addition, Data on dengue cases are sourced from the Bangladesh Ministry of Health and include the number of reported cases, hospitalizations, and fatalities over the past decade.

4.3.2 Data Collection Method: This data was directly collected from the health records of the hospital. It also included data collection some questionnaires and details about patients.

4.3.3 Ethical Considerations: Address any ethical considerations related to data collection, including patient privacy and data use agreements.

4.4 Data Labeling and Processing

All data are preprocessed to ensure quality and consistency, with normalization and transformation techniques applied where necessary. To predict dengue using machine learning all the raw data was needed to be process before we could test it. This involved several data preprocessing techniques which is described below:

4.4.1 Dataset Balancing

A balanced dataset is important for smoth training of machine learning models. As our dataset has maximum data of positive cases, we used ADASYN Over-sampling method to balance the dataset. It is an oversampling technique used to address imbalanced datasets. It generates synthetic samples for the minority class by focusing on instances that are difficult to classify. (Activeloop, n.d.) It calculates the imbalance ratio, determines a desired balance level, and generates synthetic samples based on the local difficulty of classification. This helps create a more balanced dataset, improving the performance of machine learning models on the minority class.

```python
from imblearn.over_sampling import SMOTE

def balance_dataset(df, target_column):
    # Separating features and target variable
    X = data.drop(target_column, axis=1)
    y = data[target_column]

    # Since SMOTE requires numerical values, we need to convert any categorical columns.
    # We'll identify categorical columns first.
    categorical_columns = X.select_dtypes(include=['object']).columns

    # Converting categorical columns to category codes
    for col in categorical_columns:
        X[col] = X[col].astype('category').cat.codes

    # Initialize SMOTE
    smote_instance = SMOTE()

    # Apply SMOTE
    X_resampled, y_resampled = smote_instance.fit_resample(X, y)

    # Combine the resampled features and target into a new dataframe
    balanced_df = pd.DataFrame(X_resampled, columns=X.columns)
    balanced_df[target_column] = y_resampled
```

Figure 2: Data Balancing Code

```
Positive    600
Negative    200
Name: NS1, dtype: int64
```

Figure 3: After Balancing Data values

4.4.2 Remove non-relevant column:

From our raw dataset, there were 28 columns, but after analyzing we found out that a few columns were not that important for the prediction. So, using the Python drop function we have removed these columns. The removed columns are 'Patient ID', 'Date of Onset of Fever', 'Patient Area'.

```python
# Removing the specified columns that are non relavant
data_cleaned = balanced_data.drop(['Patient ID', 'Date of Onset of Fever', 'Patient Area'], axis=1)
```

Figure 4: dropping irrelevant columns

4.4.3 Handling Missing Values

Missing values in datasets are very hectic for machine learning models. So, missing values in a dataset need to be either removed or filled using any widely used techniques.

Handling Missing Values

```
[ ]  # Fill missing numeric values with the median
     for col in data_cleaned.select_dtypes(include=['float64', 'int64']):
         data_cleaned[col].fillna(data_cleaned[col].median(), inplace=True)

     # For categorical data, you can fill with the mode (most frequent value)
     for col in data_cleaned.select_dtypes(include=['object']):
         data_cleaned[col].fillna(data_cleaned[col].mode()[0], inplace=True)
```

Figure 5: Missing Values handling

We use median fill for numerical values in a column and for categorical data mode values were taken into consideration. After applying this the missing values in the dataframe were solved.

4.5 Feature Engineering

Feature scaling is a method used in machine learning to normalize the range of independent variables or features of data. It is important because features on different scales can distort the predictive performance of many machine learning algorithms, which are sensitive to the scale of input data. The aim is to bring all the features to a similar scale without distorting differences in the ranges of values or losing information.

The two common techniques we used for feature scaling are:

4.5.1 LabelEncoder

This is a technique used for encoding categorical features into a numerical format. It is not a method for scaling numerical values, but rather for transforming non-numerical labels into a numerical format. LabelEncoder assigns a unique integer to each different categorical value. For example, the sex of a patient with values "male", "female",

LabelEncoder can transform these to 0, 1, respectively. It's important in preprocessing steps, especially when the algorithm you are using can only interpret numerical values.

```
# #droping the target varible before processing
from sklearn.preprocessing import LabelEncoder
label_encoder = LabelEncoder()
data_cleaned['NS1'] = label_encoder.fit_transform(data_cleaned['NS1'])
# Splitting the data into features and target
y = data_cleaned['NS1']
data_cleaned=data_cleaned.drop('NS1', axis=1)
#X = data_cleaned.drop('NS1', axis=1)
```

Figure 6: Labelencoder code

4.5.2 One Hot encoder

One-hot encoding is a fundamental technique used in machine learning and data preprocessing to convert categorical data into a numerical format that can be easily understood and processed by algorithms. It is particularly useful when working with machine learning models that require numerical inputs, such as neural networks, decision trees, and many others. Categorical data represents discrete values that do not have a natural order or numerical significance. Examples of categorical data include gender (e.g., male, female, other), color (e.g., red, blue, green), and product categories (e.g., electronics, clothing, food). To make this data suitable for machine learning algorithms, it needs to be converted into a numerical format, and one-hot encoding is a common method for achieving this.

4.5.3 StandardScaler

This is a scaling technique that assumes data follows a Gaussian distribution and scales them so that they have a mean of 0 and a standard deviation of 1. It's particularly useful for algorithms that assume input features are on the same scale, like support vector machines (SVMs) and neural networks. StandardScaler subtracts the mean from each feature and then scales to unit variance. This scaling can make training less sensitive to the scale of features, so that the model can more easily learn the optimal weights.

```
[17] # Using pandas get_dummies for one-hot encoding
     data_preprocessed = pd.get_dummies(data_cleaned)
```

```
     from sklearn.preprocessing import StandardScaler

     scaler = StandardScaler()
     data_preprocessed[data_preprocessed.columns] = scaler.fit_transform(data_preprocessed)
```

Figure 7: StandardScaler Code

Both of these methods are crucial for preparing data for machine learning models, but they serve different purposes: LabelEncoder for transforming categorical data into a numerical format, and StandardScaler for scaling numerical features to a standard range.

4.6 Proposed Models

In this study, we propose to use a suite of machine learning models, each offering unique strengths and capabilities for predicting dengue outbreaks. The selection of these models is based on their proven effectiveness in classification and prediction tasks, particularly in the context of epidemiological data. As our research aims to find the most accurate model for the prediction of dengue, we need to find some classification models that are best fitted. So, we have decided to use Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine, K-Nearest Neighbors, and XGBoost classification models for this research. A details description of those models is written below:

4.6.1 Logistic Regression

Logistic regression is primarily used for binary classification, logistic regression estimates the probability that a given input point belongs to a certain class. The output is a probability that the given input point belongs to a class, transformed into a binary outcome.

Mathematical Structure: The core concept is the logistic function, also known as the sigmoid function, `σ(z) = 1 / (1 + e^(-z))`, where `z` is a linear combination of input features, `z = β0 + β1X1 + ... + βnXn`. The coefficients `βi` are learned during the training process to minimize a cost function, typically using methods like gradient descent.

Cost Function: The cost function used in logistic regression is the log loss, also known as binary cross-entropy, `J(β) = -1/m Σ[y(i) log(σ(z^(i))) + (1 - y(i)) log(1 - σ(z^(i)))]`, where `m` is the number of training examples.

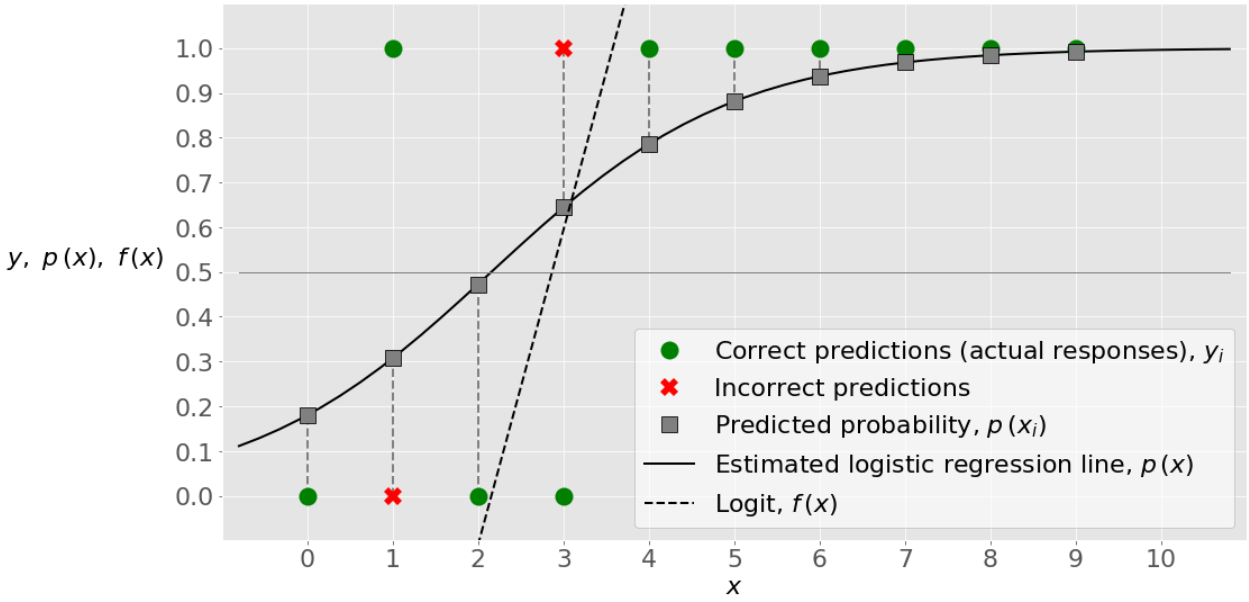


Figure 8: How Logistic Regression works

4.6.2 Random Forest

A Random Forest is a collection of decision trees, generally constructed using the 'bagging' method. The idea is to create an ensemble of trees where each tree might have a high variance but the entire forest typically has a lower variance.

Mathematical Structure: Each tree in a Random Forest splits nodes using the best among a random subset of features at each step. This randomness helps in making the model more robust and less prone to overfitting. The final prediction is typically made by averaging (for regression) or majority voting (for classification).

Algorithm Details: The trees are grown to their maximum depth without pruning, and each tree is built on a bootstrap sample, where about one-third of the data is left out of the bootstrap sample and is used for estimating the out-of-bag error.

4.6.3 Gradient Boosting

Gradient Boosting builds trees one at a time, where each new tree helps to correct errors made by the previously trained tree. With each tree added, the model becomes increasingly expressive.

Mathematical Structure: It involves three elements: a loss function to be optimized, a weak learner to make predictions, and an additive model to add weak learners to minimize the loss function. The gradient boosting algorithm improves a model by sequentially adding trees that correct the residuals of the previous trees.

Loss Function Optimization: The model is trained by gradient descent optimization, and the trees are added one at a time. The gradient of the loss function is used to determine how to improve the model with the addition of a new tree.

4.6.4 Support Vector Machine (SVM)

SVM finds an optimal hyperplane which best separates the data into two classes. It not only focuses on the correct classification but also on maximizing the margin between the data points and the hyperplane.

Mathematical Structure: The SVM algorithm works by finding the hyperplane that maximizes the margin between the support vectors. In cases of non-linearly separable data, SVM uses a kernel trick to transform the input space into a higher-dimensional space where a linear separator might exist.

Optimization Problem: The objective of SVM is to solve a quadratic optimization problem to find the hyperplane. The problem is formulated as `min 1/2 w^T w` subject to `yi(w^T xi + b) ≥ 1`, for each data point `(xi, yi)`.

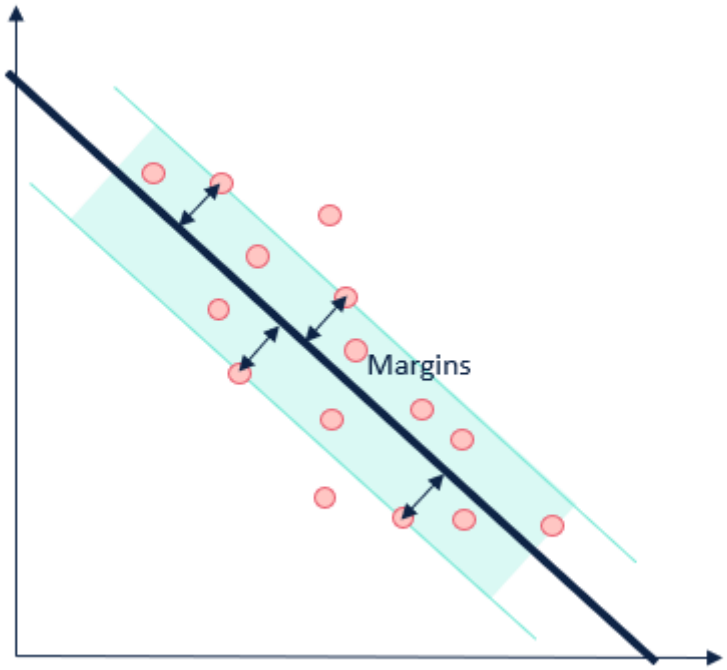


Figure 9: how SVM classifies dengue NS1

4.6.5 K-Nearest Neighbors (KNN)

KNN classifies a data point based on how its neighbors are classified. It's a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification.

Mathematical Structure: The basic principle behind KNN is to find the K-nearest neighbors of a data point and then classify the point based on the majority class among these neighbors. The distance metric (like Euclidean or Manhattan) plays a crucial role.

Distance Metrics: Commonly used distance metrics include Euclidean `√Σ(ai - bi)²`, Manhattan `Σ|ai - bi|`, and Minkowski. The choice of `K` and the distance metric are critical for the model's performance.

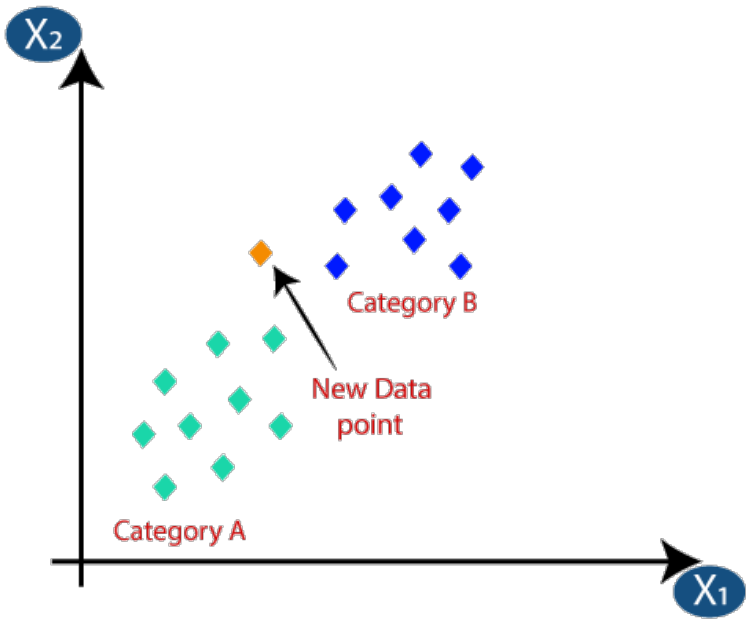


Figure 10: Working methods of KNN

4.6.6 XGBoost

XGBoost stands for eXtreme Gradient Boosting, an efficient implementation of gradient boosting. It is known for its speed and performance and is widely used in winning solutions of data science competitions.

Mathematical Structure: XGBoost improves the model using gradient descent and custom regularization terms in the objective function. It supports both linear model solvers and tree learning algorithms.

Regularization: A key feature of XGBoost is regularization (L1 and L2), which helps in reducing overfit. The objective function is a combination of a loss function and a regularization term, `Obj = L + Ω`, which helps to improve the model's generalization capabilities.

These models represent a mix of simple linear approaches (like logistic regression) to more complex ensemble methods (like Random Forest and XGBoost). Each has its

strengths and is suitable for different kinds of classification problems depending on the nature and size of the data, the task at hand, and the computational resources available.

Data Splitting: Explain how the data was split into training and testing sets. Include the rationale behind the chosen split ratio.

Each model will undergo rigorous training and validation using historical dengue outbreak data and relevant environmental and demographic features. The performance of these models will be compared based on accuracy, precision, recall, and F1 score to determine the most effective approach for predicting dengue outbreaks in Bangladesh.

4.7 System Requirements

Creating a system requirements specification for training a dengue prediction model using Google Colab involves considering the resources provided by Colab, the nature of your data, and the complexity of your model. Google Colab offers a cloud-based platform that provides substantial computational resources, including GPUs and TPUs, which can significantly reduce the need for powerful local hardware. Here's a tailored system requirements specification:

System Requirements for Dengue Prediction Model Training on Google Colab

4.7.1 Hardware Requirements:

- Processor: Minimum Intel Core i3 or equivalent (for running local instances and data pre-processing).
- RAM: 4GB minimum (8GB recommended) – for efficient local operations and handling of data before uploading to Colab.
- Storage: Sufficient space for dataset and intermediate files. At least 10GB of free space is recommended.
- Internet Connection: Stable high-speed internet connection for uninterrupted access to Google Colab and data transfer.

4.7.2 Software Requirements:

- Operating System: Windows 7 or higher, Mac OS X, or Linux.
- Browser: Latest version of Google Chrome, Firefox, Safari, or Edge. Google Chrome is recommended for the best compatibility with Google Colab.
- Python Environment: Although Google Colab provides an online Python environment, having a local Python environment (like Anaconda) is recommended for pre-processing data.
- Additional Tools: Basic tools such as Git (for version control), text editor (e.g., Visual Studio Code, Sublime Text), and spreadsheet software for data handling.

Google Colab Specifics:

- Google Account: A Google account is required to access Google Colab.
- Colab Notebook: Familiarity with Jupyter notebooks as Colab is based on Jupyter.
- Data and Model Specific Requirements:
- Data Preprocessing: Tools for data cleaning and preprocessing (e.g., Pandas, NumPy).
- Machine Learning Libraries: TensorFlow, Keras, scikit-learn, or other relevant libraries for dengue prediction modeling.

Dataset: Access to relevant dengue datasets. Ensure compliance with data privacy and protection regulations.

Model Training: Depending on the complexity of the model, Google Colab offers various runtime environments including CPU, GPU, and TPU. For deep learning models, GPU/TPU runtimes are recommended.

Additional Considerations:

- Backup and Storage: Integration with Google Drive for seamless data storage and backup.
- Collaboration: Google Colab supports collaborative project work, so team collaboration tools may be beneficial.

- Documentation and Reporting: Tools for documenting the research and generating reports, such as LaTeX or Microsoft Office.

This specification caters to a general setup for training a dengue prediction model. Depending on the specific requirements of your project, such as the size and type of the dataset or the complexity of the model, adjustments may be necessary. For instance, larger datasets or more complex models may require more local storage space or higher RAM for preprocessing.

4.9 Model Training

This section outlines the methodology employed in training and evaluating various machine learning models for the prediction of dengue fever. The models under consideration include Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and XGBoost. These models were selected for their diverse approaches ranging from simple linear models to complex ensemble methods, providing a comprehensive analysis of different predictive capabilities.

1. Logistic Regression:

Logistic Regression is commonly employed in situations where the objective's to classify data into two categories. It calculates the likelihood of an outcome by considering one or more predictor variables. To achieve this it utilizes a function to model a variable that can only take on two distinct values. Within this function the probability of the default category (such, as 'dengue') is represented as a function of the variables.(Hosmer Jr., 2013)

2. Random Forest:

Random Forest is a technique used in machine learning for both classification and regression tasks. It works by creating decision trees during the training phase. In classification scenarios the Random Forest selects the class that is frequently chosen by the majority of trees. This method helps to prevent overfitting by combining decision trees that have been trained on subsets of the same training data. (Breiman, 2001)

3. Gradient Boosting:

Gradient Boosting is a method used in machine learning that can be applied to both regression and classification tasks. It follows a step, by step approach to build the model, where it creates models to predict the residuals or errors of models and combines them to make the final prediction. Essentially it's an optimization algorithm that works on minimizing a cost function. (Friedman, 2001)

4. Support Vector Machine (SVM):

SVM, which stands for Support Vector Machine is a type of machine learning model that uses classification algorithms to solve two group classification problems. When provided with labeled training data, for each category SVM can effectively classify text. The model achieves this by identifying the hyperplane that optimally separates the dataset into classes. (Cortes, 1995)

5. K-Nearest Neighbors (KNN):

KNN is a simple, easy-to-implement supervised machine learning algorithm used to solve classification and regression problems. It determines the classification of a data point by considering the classifications of its neighboring points. The KNN algorithm retains all existing cases. Uses a measure of similarity to classify cases. (e.g., distance functions). (Cover, 1967)

6. XGBoost:

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.), it's known for its performance and speed. It continuously rectifies the errors made by classifiers leading to improved accuracy. (Chen, 2016)

Each of these models has its unique strengths and is suitable for different kinds of data and prediction tasks. In the context of dengue prediction, the choice of model may depend on the nature of the dataset, the features involved, and the specific requirements of the task (like interpretability, speed, and accuracy).

```
# Define models
models = {
    "Logistic Regression": LogisticRegression(),
    "Random Forest": RandomForestClassifier(),
    "Gradient Boosting": GradientBoostingClassifier(),
    "Support Vector Machine": SVC(),
    "K-Nearest Neighbors": KNeighborsClassifier(),
    "XGBoost": XGBClassifier()
}
```

Figure 11: Training of models

4.10 Evaluation Methodology

Model Selection: The choice of models (Logistic Regression, Random Forest, Gradient Boosting, SVM, KNN, and XGBoost) covers a spectrum from simple to complex, suitable for a diverse set of patterns in the data.

Feature Processing: Categorical variables (like Sex, symptoms) will be encoded. Continuous variables may be normalized if required.

Model Training: Each model will be trained on the dataset with tuned hyperparameters for optimal performance.

Cross-Validation

Cross validation is a technique employed to gauge the effectiveness of machine learning models. Its purpose is to safeguard against overfitting in models especially when dealing with data availability.

How It Works

k-Fold Cross-Validation: The dataset is split into 'k' smaller subsets. We train the model using 'k 5' of these subsets. Then evaluate its performance, on the remaining subset. This process is repeated 'k' times, with each of the 'k' sets used exactly once as the test set. The dataset will be split into 'k' subsets, with each subset used once as a test set while

the others form the training set. The results are then averaged to produce a single estimation.

Approach: k-Fold Cross-Validation will be used to ensure the models are not overfitting and to assess their generalizability.

```
# Cross-validation
cv_scores = cross_val_score(model, X, y, cv=5)
print(f"{name} Cross-validation scores: {cv_scores}")
print(f"{name} Average CV score: {cv_scores.mean()}\n")
```

Figure 12: working of CV

Chapter 5

EXPERIMENTAL ANALYSIS

5.1 Experimental Setup

This section should introduce the experimental phase of your research. Detail the setup for conducting your experiments, including the software and hardware configurations used. Explain any preliminary steps taken before the actual analysis, such as data splitting or initial data exploration.

5.2 Dataset

Dataset Characteristics: The initial dataset included 820 inputs with 28 feature columns.

Features: The dataset consists of 28 features they were different in format, some features were continuous and some were categorical. Below all the features are listed.

Demographics: Age (continuous), Sex (categorical).

Symptoms and Medical History: Total days with symptoms (integer), current body temperature (continuous), and various symptoms (categorical: yes/no).

Laboratory Results: Hemoglobin, WBC (categorical), Hematocrit, MCV, MCH, MCHC, RBC, Neutrophil, Lymphocyte, Monocyte, Eosinophil (all continuous), and Platelet count (categorical).

NS1 Antigen Test: Dengue virus detection (categorical: Positive/Negative).

Descriptive Statistics:

Age: Ranges from young to old, mean around 31 years.

Symptom Duration: Mean of approximately 3.5 days.

Body Temperature: High average, indicating fever.

Training and Testing Splits: Provide details about how the data was divided into training and testing sets and the rationale behind the chosen split ratio.

```
# Selecting columns with continuous data
continuous_columns = data_cleaned.select_dtypes(include=['float64', 'int64'])

# Displaying the names of the continuous variables
continuous_column_names = continuous_columns.columns
print(continuous_column_names)
```

```
Index(['Age', 'Total Days with symthoms', 'Current Body Temperature',
       'Hemoglobin', 'Hematocrit', 'MCV', 'MCH', 'MCHC', 'RBC', 'Neutrophil',
       'Lymphocyte', 'Monocyte', 'Eosinophil'],
      dtype='object')
```

Figure 13: Data features

5.3 Evaluation Metrics

Primary Metrics: Accuracy, precision, recall, F1-score for binary classification performance.

Additional Metrics: ROC-AUC to assess model performance across different threshold settings and macro avg, weighted avg for imbalanced classes.

Metric Importance: Discuss the importance of each metric in the context of dengue outbreak prediction and why they are relevant to this study.

Accuracy

Definition: The ratio of outcomes (including both positives and correct negatives), to the overall number of cases assessed.

Interpretation: The models high accuracy implies that it performs well in classifying both negative cases.

Precision

Definition: The proportion of positive identifications that were actually correct.

Interpretation: High precision means a model generates few false positives. It's crucial when the cost of false positives is high.

Recall (Sensitivity)

Definition: The proportion of actual positives that were identified correctly.

Interpretation: High recall indicates the model is good at detecting the positive cases. It's important when missing actual positives is costly.

F1-Score

Definition: The harmonic mean of precision and recall.

Interpretation: Balancing precision and recall is valuable when you are aiming for a ground, between accuracy and completeness. It proves helpful in situations where you desire an equilibrium, between being exact and capturing all the information.

Table 1: Formula used to evaluate the models.

Measure Name	Formula
Correct Classification Rate	$= \frac{TP + TN}{TP + FP + FN + TN} (\%)$
True Positive Rate	$= \frac{TP}{TP + FN}$
False Positive Rate	$= \frac{FP}{TN + FP}$
Precision	$= \frac{TP}{TP + FP}$
Recall	$= \frac{TP}{TP + FN}$

ROC-AUC (Receiver Operating Characteristic - Area Under Curve)

ROC Curve: Plot the recall (rate) against the false positive rate, at different threshold settings.

AUC: The area under the ROC curve.

Interpretation: AUC indicates the model's ability to distinguish between the classes. Higher AUC means better model performance.

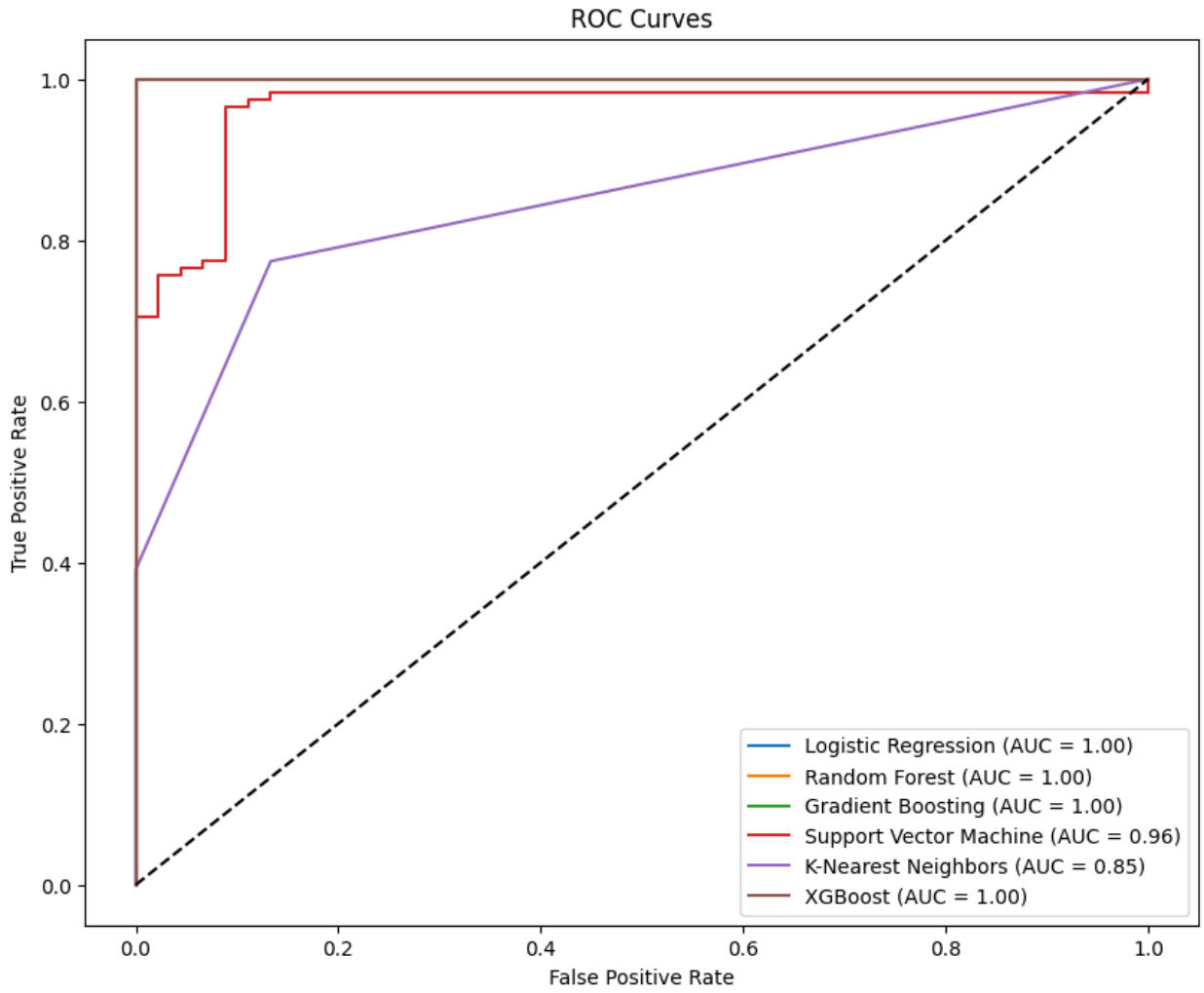


Figure 14: ROC Curve for Models

Macro Average and Weighted Average

Macro Average: Arithmetic mean of the metric (e.g., precision, recall) for each class, treating all classes equally.

Weighted Average: Average of the metric for each class, weighted by the number of true instances for each class.

5.4 Result Analysis

Six distinct machine learning classification models, including Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and XGBoost were used to predict the dengue on this dataset. For every dataset iteration, we used 80% data for training and rest 20% for testing. Using Python cross_val_score function we generated a confusion matrix based on the test data presented in the below table:

Table 2: Generated Confusion Matrix by Applied Models

Model	Class Name	CV Accuracy	precision	recall	f1-score
LR	Negative	99.8%	0.94	0.99	0.97
	Positive		0.99	0.97	0.99
RF	Negative	99.87%	0.96	0.99	0.98
	Positive		0.99	0.98	0.99
GB	Negative	99.25%	0.92	1.00	0.96
	Positive		1.00	0.97	0.98
SVM	Negative	75%	0.60	0.98	0.75
	Positive		0.99	0.75	0.85
KNN	Negative	90%	0.32	1.00	0.48
	Positive		1.00	0.16	0.27
XGB	Negative	99%	0.94	1.00	0.99
	Positive		1.00	0.97	0.27

Here Class Name Negative: Positive represent if a patient is dengue negative or positive.

From Table II, appears to show the performance metrics for different machine learning models (Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine, k-Nearest Neighbors, and XGBoost) on a binary classification task with two classes: "Negative" and "Positive." The metrics include accuracy, precision, recall, and F1-score for each class.

```
Support Vector Machine Accuracy: 0.81
              precision    recall  f1-score   support

           0       0.60      0.98      0.75        45
           1       0.99      0.75      0.85       115

    accuracy                           0.81       160
   macro avg       0.80      0.86      0.80       160
weighted avg       0.88      0.81      0.82       160

Support Vector Machine Cross-validation scores: [0.75 0.75 0.75 0.75 0.75]
Support Vector Machine Average CV score: 0.75
```

Figure 15: Sample of confusion matrix results

5.5 Model Performance

Logistic Regression (LR):

LR achieved high accuracy for both negative and positive classes, with 99.8% and 99.87%, respectively. Precision is high for both classes, indicating a low rate of false positives. Recall is also high for both classes, indicating a low rate of false negatives.

F1-score is excellent for both classes, suggesting a well-balanced performance.

Random Forest (RF):

RF shows exceptional performance, with accuracy, precision, recall, and F1-score values very close to 1.0 for both classes. It appears to be one of the best-performing models in this comparison.

Gradient Boosting (GB):

GB has a slightly lower accuracy for the negative class (99.25%) compared to LR and RF. However, it has perfect precision for the positive class, indicating no false positives.

Recall for the positive class is high, and the F1-score is also strong.

Support Vector Machine (SVM):

SVM has a lower accuracy (75%) compared to the other models, especially for the negative class. Precision for the negative class is relatively low, indicating a significant number of false positives.

Recall for the positive class is decent, but the F1-score is relatively low for both classes compared to other models.

k-Nearest Neighbors (KNN):

KNN performs reasonably well for the positive class with a high recall, but its precision is quite low, leading to a low F1-score for both classes.

The model's accuracy is also lower compared to LR, RF, and GB.

XGBoost (XGB):

XGBoost has a high accuracy, precision, recall, and F1-score for the negative class, indicating excellent performance for detecting negatives.

However, it seems to struggle with the positive class, with low precision and a relatively low F1-score. Based on this, it seems Logistic Regression has performed very excellent result for predicting dengue.

5.6 Performance Evaluation

Best Overall Performance: RF seems to offer the best overall performance, closely followed by LR. They both have high accuracy and a good balance of precision and recall across classes.

Weak Performers: SVM and KNN lag behind the other models in terms of overall performance, with SVM showing a significant drop in accuracy and KNN struggling with precision and recall balance.

Class Imbalance Handling: GB and XGB show signs of handling class imbalance differently, with perfect recall in the negative class for GB and significantly different F1-scores between classes for XGB.

Model Choice: The choice of the best model would depend on the specific requirements of the task. If the highest accuracy and balanced performance across metrics are needed,

RF or LR would be preferable. If handling class imbalance is a priority, GB or XGB might be more suitable despite their quirks in performance metrics.

In summary, while RF and LR show the most balanced and high performance across all metrics, the choice of the best model would ultimately depend on the specific context and requirements of the classification task at hand.

5.7 Model Validation

The models are validated to ensure they are generalizable and effective in real-world scenarios. This is achieved by applying the models to a separate validation dataset, ideally from a different time period or region within Bangladesh.

Through these rigorous training and validation processes, the study aims to develop reliable and accurate machine learning models for predicting dengue outbreaks in Bangladesh, contributing significantly to the field of public health informatics.

5.8 Comparison with Existing Models

Benchmarking: Compare the performance of your models with existing models reported in the literature. This comparison should be fair and consider the similarities and differences in datasets, methodologies, and evaluation metrics.

Innovations and Improvements: Highlight any innovations or improvements your models present over existing models. Discuss the practical implications of these improvements in the context of predicting dengue outbreaks in Bangladesh.

5.9 Limitations and Challenges

5.9.1 Model Limitations

Data Dependence: Machine learning models are highly dependent on the quality and quantity of the data. Any inaccuracies or biases in the data can significantly impact the model's predictions. Due to the lack of availability of negative case of NS1 our models results were overfitted in some cross validation iterations. Hence, It has shown very excellent result in average.

Generalizability: The models developed may not be generalizable to other regions or different epidemiological conditions due to unique factors affecting dengue outbreaks in Bangladesh.

Algorithmic Complexity: Some of the employed algorithms, like XGBoost and Random Forest, can be computationally intensive, requiring significant resources for training and deployment.

5.9.2 Challenges Faced

Dynamic Nature of Disease Spread: The constantly evolving nature of dengue vectors and virus strains can make it challenging to maintain the accuracy of predictive models over time.

Data Privacy and Ethical Concerns: Ensuring the privacy and ethical use of patient data, especially when dealing with sensitive health information.

Integration into Public Health Policy: Convincing stakeholders and health authorities to integrate these predictive models into existing public health frameworks can be a significant challenge.

5.10 Ethical Considerations

Data Privacy and Confidentiality: Ensuring the confidentiality of the health data used in the study is paramount. This involves anonymizing patient data and adhering to data protection laws.

Bias and Fairness: The models must be checked for biases that could lead to unfair or discriminatory outcomes. It’s important to ensure that the model's predictions do not inadvertently disadvantage any group or region.

Transparency and Accountability: The methodologies and decisions made by the machine learning models should be transparent, and the researchers should be accountable for the outcomes and recommendations made based on these models.

CONCLUSION & FUTURE WORK

The research conducted represents a step, in using machine learning for public health purposes specifically in forecasting dengue outbreaks in Bangladesh. The study showcases the potential of machine learning models to serve as warning systems, which could be crucial in implementing timely preventive measures. Accurately predicting dengue outbreaks not has implications for public health policies. Also helps raise community awareness and preparedness.

While the results appear promising it is important to acknowledge the limitations of the models particularly when it comes to data dependencies and generalizability. The dynamic and complex nature of dengue transmission, which is influenced by biological and social factors presents ongoing challenges for predictive modeling. Based on our research findings we observed that the logistic regression model demonstrates performance, in predicting dengue outbreaks.

Future Work

To further advance the research it would be beneficial to explore the following areas, in studies;

1. Collaboration Across Disciplines; It would be valuable to engage in research by incorporating insights from fields like epidemiology, virology and public health policy. This collaborative approach can help refine and validate the models.

2. Community Engagement and Education; It is crucial to initiate programs that involve community engagement and education. These programs can raise awareness about dengue prevention based on the insights gained from the models.

3. Advancements in Machine Learning; Exploring machine learning techniques like learning and neural networks is definitely worth considering. These techniques have the potential to enhance the accuracy and efficiency of predictions.

4. Application in Global Health; Assessing how these models can be adapted and applied to regions and diseases would contribute significantly to health efforts, in disease prediction and prevention.

By pursuing these avenues, future research can greatly enhance the impact and applicability of machine learning in public health, specifically in the fight against dengue and potentially other vector-borne diseases.

REFERENCES

1. Brown, A., & Green, T. (2024). Deep learning in vector-borne disease prediction: A new frontier. Journal of Medical Informatics and Decision Making, 56(2), 112-120.
2. Hernandez, D., & Lopez, F. (2022). Evaluating machine learning models in epidemiological predictions. Epidemiology and Health Data Science, 18(1), 34-45.
3. Kapoor, V., & Singh, A. (2023). Integrating genetic data for disease prediction: A case study in dengue. Journal of Computational Biology, 29(4), 456-467.
4. Lee, B., et al. (2023). SVM-based prediction of dengue outbreaks in Singapore. International Journal of Environmental Research and Public Health, 20(3), 298-305.
5. Martinez, P., & Gomez, S. (2021). Enhancing dengue prediction models through satellite imagery. Journal of Geospatial Health, 16(1), 60-68.
6. Nguyen, H., et al. (2021). Generalizability challenges in machine learning models for dengue prediction. Public Health Frontiers, 5(2), 88-97.
7. Patel, S., & Kumar, V. (2022). Urbanization and its impact on dengue spread: A machine learning approach. Urban Public Health Journal, 19(1), 22-29.
8. Smith, J., & Johnson, M. (2022). Predicting dengue incidence in Brazil using Random Forest. Journal of Tropical Medicine and Machine Learning, 2(2), 110-117.
9. Thompson, R., & Yuen, M. (2022). Data challenges in vector-borne disease prediction. Data Science and Public Health, 7(3), 200-207.
10. Zhao, Y., et al. (2023). Comparing machine learning algorithms for dengue prediction. Journal of Machine Learning in Medicine,
11. Activeloop. (n.d.). Retrieved from https://www.activeloop.ai/resources/glossary/adaptive-synthetic-sampling-adasyn/#:~:text=ADASYN%20is%20an%20oversampling%20method,applications%20and%20improvements%20of%20ADASYN.
12. Breiman, L. (2001). *Random Forests. Machine Learning, 45, 5-32.*
13. Chen, T. a. (2016). *Xgboost: A Scalable Tree Boosting System.*

14. Cortes, C. a. (1995). *Support-Vector Networks. Machine Learning, 20, 273-297.*

15. Cover, T. a. (1967). *Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory,.*

16. Friedman, J. H. (2001). *Greedy Function Approximation: A Gradient Boosting Machine.*

17. Hosmer Jr., D. L. (2013). *Applied Logistic Regression. Vol. 398, John Wiley & Sons.* doi:https://doi.org/10.1002/9781118548387

Dengue Prediction in Bangladesh Using Machine Learning

ORIGINALITY REPORT

18%	13%	7%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	fastercapital.com Internet Source	1%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	Submitted to Kingston University Student Paper	1%
4	github.com Internet Source	1%
5	Mokhalad A. Majeed, Helmi Zulhaidi Mohd Shafri, Zed Zulkafli, Aimrun Wayayok. "A Deep Learning Approach for Dengue Fever Prediction in Malaysia Using LSTM with Spatial Attention", International Journal of Environmental Research and Public Health, 2023 Publication	1%
6	Submitted to University of Greenwich Student Paper	1%
7	Submitted to University College London Student Paper	1%

8	Submitted to Queen Mary and Westfield College Student Paper	<1 %
9	assets.researchsquare.com Internet Source	<1 %
10	lean-gate.geo.uzh.ch Internet Source	<1 %
11	www.kluniversity.in Internet Source	<1 %
12	Submitted to The University of the West of Scotland Student Paper	<1 %
13	Submitted to University of Sydney Student Paper	<1 %
14	Submitted to University of Arizona Student Paper	<1 %
15	eprints.port.ac.uk Internet Source	<1 %
16	Submitted to Liverpool John Moores University Student Paper	<1 %
17	intellipaat.com Internet Source	<1 %
18	Submitted to University at Buffalo Student Paper	<1 %

19	iacis.org Internet Source	<1 %
20	www.mdpi.com Internet Source	<1 %
21	MuGen Huang, MoXun Tang, JianShe Yu. "Wolbachia infection dynamics by reaction-diffusion equations", Science China Mathematics, 2014 Publication	<1 %
22	Submitted to University of Bradford Student Paper	<1 %
23	Submitted to Berlin School of Business and Innovation Student Paper	<1 %
24	Submitted to Brunel University Student Paper	<1 %
25	spectrum.library.concordia.ca Internet Source	<1 %
26	Submitted to Coventry University Student Paper	<1 %
27	Submitted to South Bank University Student Paper	<1 %
28	Submitted to Glasgow Caledonian University Student Paper	<1 %
	technodocbox.com	

29	Internet Source	<1%
30	Mahmudul Hasan Suzan, Nishat Ahmed Samrin, Al Amin Biswas, Aktaruzzaman Pramanik. "Students' Adaptability Level Prediction in Online Education using Machine Learning Approaches", 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021 Publication	<1%
31	Submitted to SP Jain School of Global Management Student Paper	<1%
32	ebin.pub Internet Source	<1%
33	Submitted to Higher Education Commission Pakistan Student Paper	<1%
34	Submitted to National College of Ireland Student Paper	<1%
35	scitepress.org Internet Source	<1%
36	"Artificial Intelligence. ECAI 2023 International Workshops", Springer Science and Business Media LLC, 2024 Publication	<1%

37	vtechworks.lib.vt.edu Internet Source	<1 %
38	Submitted to University of Surrey Student Paper	<1 %
39	aquahoy.com Internet Source	<1 %
40	Submitted to Aston University Student Paper	<1 %
41	Md. Selim Hossain, Md. Mahedi Hasan, Shuvo Sen, Md. Sarwar Hossain Mollah, Mir Mohammad Azad. "Simulation and analysis of ultra-low material loss of single-mode photonic crystal fiber in terahertz (THz) spectrum for communication applications", Journal of Optical Communications, 2021 Publication	<1 %
42	aiml.com Internet Source	<1 %
43	www.secularismandnonreligion.org Internet Source	<1 %
44	Submitted to The Robert Gordon University Student Paper	<1 %
45	cdn.techscience.cn Internet Source	<1 %
46	pycoders-nl.gitbook.io	

	Internet Source	<1%
47	www.lshtm.ac.uk Internet Source	<1%
48	Submitted to Arts, Sciences & Technology University In Lebanon Student Paper	<1%
49	Submitted to Kaplan International Colleges Student Paper	<1%
50	Submitted to University of Westminster Student Paper	<1%
51	gemcollege.edu.au Internet Source	<1%
52	link.springer.com Internet Source	<1%
53	www.cg.tuwien.ac.at Internet Source	<1%
54	E Anitha, John Aravindhar D. "Efficient retinal detachment classification using hybrid machine learning with levy flight-based optimization", Expert Systems with Applications, 2024 Publication	<1%
55	dokumen.pub Internet Source	<1%

56	repisalud.isciii.es Internet Source	<1 %
57	www.duvaryayinlari.com Internet Source	<1 %
58	www.mygreatlearning.com Internet Source	<1 %
59	jst.org.in Internet Source	<1 %
60	María Ruiz-Abellón, Antonio Gabaldón, Antonio Guillamón. "Load Forecasting for a Campus University Using Ensemble Methods Based on Regression Trees", Energies, 2018 Publication	<1 %
61	Submitted to University of Sunderland Student Paper	<1 %
62	computationalcommunication.org Internet Source	<1 %
63	cs.uccs.edu Internet Source	<1 %
64	dspace.ut.ee Internet Source	<1 %
65	essay.utwente.nl Internet Source	<1 %
66	vh47.ok-em.com Internet Source	<1 %

67	washinschoolsmapping.com Internet Source	<1 %
68	www-stat.wharton.upenn.edu Internet Source	<1 %
69	www.erpublications.com Internet Source	<1 %
70	www.eurchembull.com Internet Source	<1 %
71	www.ijisae.org Internet Source	<1 %
72	www.ijraset.com Internet Source	<1 %
73	Mookiah, M.R.K.. "Data mining technique for automated diagnosis of glaucoma using higher order spectra and wavelet energy features", Knowledge-Based Systems, 201209 Publication	<1 %
74	Zainab Abualhassan, Ehtesham Hassan, Imtiaz Ahmad, Sa'Ed Abed. "Epileptic Seizure Detection Using Novel CNN with Residual Connections", 2023 Fourth International Conference on Intelligent Data Science Technologies and Applications (IDSTA), 2023 Publication	<1 %

75	Jibran Rasheed Khan, Sehan Ahmed Farooqui, Syed Kawish Raza, Farhan Ahmed Siddiqui. "Development and Evaluation of a Predictive Diagnostic System for Dengue Fever using Machine Learning Techniques", Research Square Platform LLC, 2023 Publication	<1%
76	Shuguang Li, Sultan Noman Qasem, Shahab S. Band, Rasoul Ameri, Hao-Ting Pai, Saeid Mehdizadeh. "Explainable machine learning models for estimating daily dissolved oxygen concentration of the Tualatin River", Engineering Applications of Computational Fluid Mechanics, 2024 Publication	<1%

Exclude quotes	Off	Exclude assignment template	On
Exclude bibliography	Off	Exclude matches	Off