# A MACHINE LEARNING APPROACH: PREDICTING THE NUMBER OF DENGUE PATIENTS IN BANGLADESH BASED ON CLIMATE CHANGING DATA

**BY**

**JENY AKTER**
**ID: 201-15-14253**

## FINAL YEAR DESIGN PROJECT REPORT

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

Supervised By

**Ms. Sharmin Akter (SNA)**
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

**Nazmun Nessa Moon (NNM)**
Associate Lecturer
Department of CSE
Daffodil International University

**DAFFODIL INTERNATIONAL UNIVERSITY**

**DHAKA, BANGLADESH**

**July 2024**

## APPROVAL

This Project titled **"A Machine Learning Approach : Predicting the Number of Dengue Patients in Bangladesh Based on Climate Changing Data"**, submitted by **Jeny Akter, Student ID: 201-15-14253** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 14 July 2024.

## BOARD OF EXAMINERS

**Dr. S.M Aminul Haque (SMAH)** — **Board Chairman**
Professor & Associate Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

**Md. Abbas Ali Khan (AAK)** — **Internal Examiner 1**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

**Mr. Md. Aynul Hasan Nahid (AHN)** — **Internal Examiner 2**
Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

14.07.2024

**Dr. Abu Sayed Md. Mostafizur Rahaman (ASMR)** — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

# DECLARATION

I hereby declare that this project has been done by us under the supervision of **Ms. Sharmin Akter , Senior Lecturer, Department of CSE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

**Supervised by:**

11.07.24

______________________

**Ms. Sharmin Akter**
Senior Lecturer
Department of CSE
Daffodil International University

**Co-Supervised by:**

______________________

**Nazmun Nessa Moon**
Associate Lecturer
Department of CSE
Daffodil International University

**Submitted by:**

Jeny

______________________

**Jeny Akter**
ID: 201-15-14253
Department of CSE
Daffodil International University

# ACKNOWLEDGEMENT

First I express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

I really grateful and wish our profound our indebtedness to **Ms. Sharmin Akter** , **Senior Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of IoT, Data Mining,Block Chain,NLP,Machine Learning to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

I would like to express our heartiest gratitude to **Ms. Sharmin Akter, Nazmun Nessa Moon,** and **Dr. Sheak Rashed Haider Noori**, Professor & Head**,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of our parents.

# ABSTRACT

Aedes aegypti mosquitoes are the vector of dengue fever, which can cause fatalities. Symptoms range from mild flu-like symptoms to severe diseases like shock syndrome and dengue hemorrhagic fever. In light of climate-related data, this work presents a reliable machine learning model for estimating the number of Dengue patients in Bangladesh. Various models, such as regression models and KNN, Decision Tree, and Random Forest for classification, are used in supervised learning with labeled data. The impact of dengue in 2022–2023 emphasizes how urgent it is to address this health catastrophe, as an increasing number of cases and deaths are being caused by collective neglect. In spite of previous studies, this analysis provides insightful information based on the most recent data relevant to Bangladesh. Interestingly, Random Forest performed quite well in both regression and classification, whereas Decision Tree was very effective in the former, showing an excellent f1 score of 0.85 and a marginally lower accuracy of 84.79% in comparison to Random Forest's 86.20%.

## TABLE OF CONTENTS

## LIST OF FIGURES



## LIST OF TABLES

# CHAPTER 1

# INTRODUCTION

## 1.1 Introduction

Dengue fever, a viral disease transmitted by Aedes aegypti mosquitoes, exhibits a spectrum of symptoms from mild flu-like manifestations to severe conditions such as dengue hemorrhagic fever and shock syndrome, potentially leading to fatality. The virus comprises four serotypes (DENV-1 to DENV-4), and acquiring immunity to one does not confer protection against the others. The initial stage, characterized by elevated fever, intense headaches, joint pain, nausea, vomiting, and a distinctive rash, represents a crucial phase. While traditionally associated with tropical areas, dengue can also manifest elsewhere, especially during the rainy season when mosquito populations reach their peak.

Climate change is escalating the threat of dengue fever by creating favorable conditions for Aedes aegypti mosquitoes, the primary vectors. Warmer temperatures and increased rainfall boost mosquito populations, leading to more frequent and intense outbreaks. This surge poses dire consequences, with heightened risks for vulnerable populations and significant economic burdens. Addressing the climate crisis is crucial, requiring a combination of emission reduction and effective mosquito control measures to ensure a safer future worldwide. Moreover, there exists a direct relationship between fluctuations in climate and the occurrence of dengue outbreaks [4].

Climate change is already impacting the rapid spread of dengue fever, a mosquito-borne disease, with a particular focus on countries like Bangladesh, which frequently experience dengue outbreaks. Predicting dengue cases is vital, given its significance as a major health concern in tropical regions, and climate variables, especially temperature, humidity, Rainfall, and Dengue cases play a significant role in disease transmission [2,8]. Dengue fever is a major public health problem in Asia, causing 70% of all diseases worldwide. It affects an estimated 100 million symptomatic and 300 million asymptomatic people annually, a 400% increase in recent years [7].

In Bangladesh, dengue outbreaks are worsened by urbanization, climate change, and longer rains, causing more deaths. The return of dengue virus 4 in 2022, especially in Dhaka and Rohingya camps, and COVID-19 hitting at the same time [22]. Climate change is exacerbating the dengue crisis in Bangladesh. Alterations in weather patterns, including rising temperatures and erratic rainfall, are fostering an environment conducive to the proliferation of Aedes mosquitoes, the vectors of dengue virus [1,3]. Additionally, This underscores the urgency of developing comprehensive public health strategies to mitigate the escalating dengue epidemic. As of September 16, 2023, Bangladesh has recorded over 167,700 dengue cases and 822 fatalities, with Dhaka bearing the brunt of the outbreak, which has now spread to all 64 districts.

This research is a pioneer in the development of a strong machine learning model that can estimate dengue patient numbers based on a thorough examination of climate-related data, in response to the growing threat that dengue poses to public health in Bangladesh. By using a detailed geospatial analysis, this project aims to identify cities that are disproportionately susceptible to dengue outbreaks. Secondly, it provides Bangladeshi public health authorities with an evidence-based tool for resource allocation and targeted interventions, which will improve their preparedness and response tactics. Millions of lives could be saved by this timely project, which has the potential to significantly reduce Bangladesh's dengue outbreak burden.

## 1.2 Motivation

Our study is to mitigate the health risk caused by climatic complexity by utilizing machine learning to forecast incidence and improve Dengue outbreak management in Bangladesh.The following summarizes the main arguments in favor of our study:

- **Improved Forecasting Precision:** Compared to traditional models, machine learning is able to identify complex relationships between climate change and disease, which improves readiness and forecast accuracy.

- **Real-time Dengue Evaluation:** By utilizing machine learning, dengue patterns may be tracked in real-time, facilitating early warnings for timely public health actions and averting potential outbreaks.

- **Identification of Key Climate Drivers:** For revealing significant correlations between climatic factors and dengue cases, machine learning (ML) analysis can pinpoint the main causes of epidemics. This information guides specific activities that target the most important variables, including vector control or awareness campaigns.
- **Adapting to Climate Change:** Methods of machine learning (ML) evaluate the effects of climate change on the dynamics of transmission of dengue fever. This makes it possible to adapt preventative and control techniques proactively in order to accommodate changing weather patterns.
- **Improving Resource Allocation:** Accurate projections optimize the allocation of medical resources by concentrating them in areas with the highest anticipated patient counts. This increases the efficacy and effectiveness of health care responses.
- **Potential Benefits:**

  i. Precise forecasts provide prompt planning, which lowers the number of Dengue cases and deaths.

ii. Specific public health initiatives are informed by study results to avoid dengue.
iii. In regions facing comparable difficulties, the flexible strategy helps combat dengue worldwide.

The critical requirement to address Bangladesh's rising dengue epidemic, which is being made worse by climate change, motivates our research. Our objective is to enable communities and public health authorities to confront dengue outbreaks proactively with a machine learning-based prediction model, thereby saving lives and lowering healthcare expenses.

### 1.3 Rationale of the Study

The importance and possible impact of addressing the research objectives outlined in our paper serve as the foundation for the rationale for our investigation. In this regard, we provide a comprehensive justification for the research.:

- **Tackling a serious public health issue:** Bangladesh has a serious public health threat from dengue fever, which frequently has high rates of morbidity and mortality. Medical professionals can save lives and save money on medical expenses by implementing efficient preventative measures when future epidemics are accurately predicted.
- **Addressing a knowledge gap:** Bangladesh intends to look into particular mechanisms and create a precise, specific to the setting model using climatic data to predict future dengue cases, even if previous research has linked outbreaks of dengue to climate change.
- **Establishing use of machine learning potential:** Complex interactions between dengue outbreaks and climate variables are sometimes difficult for traditional statistical models to grasp. Large datasets may teach machine learning algorithms complex patterns, which opens up a world of possibilities for more precise and nuanced predictions.

- **Tracking in real time and early warning:** With regular data updates, our machine learning model will be able to track dengue trends in real-time and provide early alerts for possible outbreaks. This confers authority onto public health authorities to execute prompt actions and alleviate the consequences of epidemics.
- **Creating interventions to particular drivers:** Our approach is able to pinpoint the primary causes of dengue outbreaks in certain areas by examining the correlations between various climate variables and dengue cases. This information can help focus targeted activities on the most influential factors, including public awareness campaigns or vector control measures.
- **The adaptation to climate change:** It is anticipated that climate change would further modify weather patterns, which could impact the dynamics of dengue transmission. Our model can be used to predict how these modifications would affect dengue epidemics in the future, enabling proactive modifications to control and preventative tactics.
- **Allocating resources optimally:** Predicting dengue cases accurately can aid in this process by allocating medical resources, such as staff, supplies, The efficacy and efficiency of public health interventions may increase as a result.

In conclusion, the dire need to treat Bangladesh's rising dengue epidemic which is being made worse by climate change is what motivates our study. Our goal is to create a resilient and flexible model for dengue case prediction by machine learning. This will allow for preventive public health measures, efficient resource management, and eventually, life-saving results.

### 1.4 Research Questions

Our research subjects of inquiry provide as a distinct focus for our investigation and its direction. In order to address the goals of our study, we have developed the following research questions:

- Can the number of dengue patients in Bangladesh be accurately predicted by a machine learning model using data on climate change?
- What differences in performance do machine learning algorithms (DecisionTree, RandomForest, LogisticRegression, etc) have when it comes to using climatic data to forecast dengue incidence in Bangladesh?
- Which combination of meteorological factors (temperature, precipitation, and humidity) enables machine learning algorithms to forecast dengue cases with the highest degree of accuracy?
- What geographical differences in the performance of the machine learning model are there within Bangladesh?
- Could the machine learning model forecast dengue epidemics in a time frame that allows for preparation and preventive measures to be taken?
- What are the main obstacles and constraints to dengue prediction in Bangladesh using machine learning models?
- Is it possible to create focused vector control or public awareness programs using the established climate factors of dengue outbreaks?

Our goal is to improve understanding of the utility, potential, and consequences of these models by investigating the effectiveness of machine learning in predicting Dengue incidence in Bangladesh through the use of climate data.

### 1.5 Expected Output

The following projected outcomes are based on our investigation and will be included in our research paper:

- **Improved Predictive Precision:** Use a range of machine learning techniques (Random Forest, Gradient Boosting, Decision Tree, etc.) to determine which model performs best at predicting dengue cases. Develop and incorporate significant climate elements beyond traditional statistical measurements in order to fully reflect the intricate relationship with dengue epidemics.
- **Real-time Surveillance and Outbreak Notifications:** Construct a real-time surveillance system by consistently gathering information from public health and meteorological databases. Forecasts that are particular to a given region should be provided for resource allocation and focused actions in various parts of Bangladesh.
- **Impact on Public Health and Policy Implications:** Develop evidence-based approaches for dengue control and climate adaptation, execute focused awareness efforts, and allocate resources for healthcare need most effectively by utilizing model projections.
- **Scalability and Continuous Improvement:** Make the model and data available to the public for wider adaption. Create a feedback loop for continuous improvement, which will guarantee flexibility and increased accuracy over time.
- **Understanding Climate-Dengue Dynamics:** Determine environmental variables and use the information to inform focused campaigns and interventions by utilizing comprehensible machine learning. For important knowledge and advancement, evaluate how future climate scenarios would affect the spread of dengue.

The results of our research demonstrate that the machine learning model can accurately forecast instances of Dengue in Bangladesh, improving health care preparedness, expanding scientific knowledge, and influencing policy responses to Dengue and climate change.

### 1.6 Report Layout

The paper explains facts and viewpoints in an organized manner. The following describes its division into different chapters and sections:

- Chapter 1: Introduction

  The first section includes an outline of the study issue along with the objective, justification, research questions, expected outcomes, and a brief description of the report structure.

- Chapter 2: Literature Review

  This chapter examines the research on dengue prediction using climate data and machine learning, with an emphasis on approaches, data sources, model performance, shortcomings, and enhancements. The presentation covers machine learning techniques for predicting dengue fever and examines relevant research in the domains of climate change and dengue prediction. A summary and comparative analysis are then given. In addition, the chapter explains the challenges the endeavor poses and expands on the issue's scope.

- Chapter 3: Methodology / Requirements Analysis & Design Specification

  This chapter describes the study methodology that was applied to the research. Along with the methods for data collection, dataset use, data pretreatment stages, statistical analysis techniques, and the recommended methodology for machine learning-based models, the study topic and tools are discussed. The chapter also discusses the hardware and software requirements needed to implement the solution.

- Chapter 4: Implementation

  The dengue prediction model's implementation is covered in detail in this chapter, along with issues with data preprocessing, model selection, training, validation, and problems.

- Chapter 5: Experimental Results & Discussion

This chapter defines the exploratory framework in order to provide an explanation of the specifics of the conducted studies. A thorough analysis and discussion are held regarding the acquired results. The research is focused on two different areas: regression analysis for patient prediction and categorization analysis. In the following discussion, significant discoveries and their consequences are emphasized along with explanations and analyses of the data.

- Chapter 6: Impact on Society, Environment, and Sustainability

The societal, environmental, and ethical ramifications of the research are examined in this chapter. The study examines the possible results of the produced models and talks about the sustainability plan created for the proposed resolutions.

Chapter 7: Summary, Conclusions, and Future Work

The last chapter provides an overview of the entire study, restating the key findings and highlighting additional implications from the conducted research. Additionally, it examines the implications of the findings and makes recommendations for future directions for investigation.

- References

This section creates an exhaustive list of all the references that are cited in the article while using a certain citation style.

- Appendix A

This appendix outlines how the course outcomes are aligned with the complex engineering problems (EP) and activities (EA) tackled in the project on "Dengue Outbreak Prediction Using Machine Learning Models."

- Plagiarism Report

  The purpose of the plagiarism report is to verify the originality and integrity of the research through a thorough analysis that compares the paper's content with existing literature to confirm that no instances of plagiarism have been detected.

The report's well-organized structure ensures that the research is presented in a methodical manner, making it easy for readers to navigate and promoting a thorough comprehension of the methodology, conclusions, and consequences of the study.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 Overview

This chapter provides an extensive review of the literature related to dengue prediction using climate data and machine learning models. Dengue fever, caused by the dengue virus and transmitted by Aedes mosquitoes, presents significant public health challenges in tropical and subtropical regions. Predicting dengue outbreaks is crucial for timely interventions and resource allocation. This review covers previous studies on dengue prediction, focusing on methodologies, data sources, machine learning models employed, and their performance. By understanding the strengths and limitations of existing works, this study aims to build on past research and propose enhanced predictive models. The chapter also identifies gaps in current knowledge and discusses open issues that need to be addressed to improve the accuracy and applicability of dengue prediction models.

## 2.2 Related Works

This section thoroughly examines previous studies on the subject, evaluating its approaches, benefits and limitations. This contextual analysis clarifies the ways in which the current study fills in knowledge gaps and enhances the field of study.

### 2.2.1 Climate Change and Dengue Prediction

The establishment of different methods and processes for identifying Dengue cases has been made easier by the integration of data on climate change with a variety of data sources. Hossain et al. [8] Dengue is primarily spread by temperature, humidity, and wind speed; on the other hand, higher rainfall has been linked to a lower number of cases of dengue.Epidemics of dengue fever can put a pressure on healthcare systems since they are unpredictable. Climate-based models can aid in the prediction of epidemics, enabling public health managers to make resource allocation and preventative decisions. When

Islam et al. [21] looked into the relationship between weather and Dengue outbreaks in Bangladesh, they discovered that there was a strong positive correlation with dew point, relative humidity, rainfall, and wind speed but no significant relationship with temperature, surface pressure, or wind speed. This finding highlighted the difficulty of developing future preventative measures.Research from Noumea, New Caledonia, Singapore, and French Guiana shown how well climate-based models forecast dengue outbreaks. The sensitivity and specificity of the models were highlighted by the Noumea study [20], which showed excellent accuracy in differentiating between epidemic and non-epidemic years. A time series Poisson regression model was used to evaluate the relationship between increased dengue cases in Singapore [19] and elevated temperature and precipitation. The French Guiana study [18] demonstrated the promise of climate-based approaches in dengue fever prediction by effectively using oceanic and atmospheric variables to forecast 80% of outbreaks and 15% of non-epidemic years. A reduction in Bangladesh's capacity for dengue transmission as a result of climate change was used by Paul et al. [9]. In a study done in Sri Lanka, Faruk et al. [5] discovered that the northeast monsoon season, humidity, precipitation, air pressure, and air pressure all have a major impact on dengue transmission. Policymakers can use these findings to improve dengue control methods. Pinto et al. [16] used principal component analysis and a Poisson regression model to study dengue case variations. The results of this study showed that the observed considerable rise in dengue cases was primarily predicted by both maximum and minimum temperatures. In a similar vein, a Singaporean study [19] used a time-series Poisson regression model to examine the connection between dengue cases and weather conditions. They discovered a direct correlation between higher precipitation and warmth and a rise in dengue cases.

### 2.2.2 Machine Learning models for dengue prediction

Dengue outbreaks based on climate change data have been predicted using a range of methods and approaches that draw from multiple data sources.The effect of weather on dengue transmission in Dhaka, Bangladesh was studied by Banu et al. [2] using a distributed lag model and a mixed Poisson time series, along with Spearman's correlation. The results showed 89% overall agreement, 84% sensitivity, and 91% specificity.For each unit increase in Maxtemp_3, the incidence rate ratio dropped by 0.759%, according to research by Tuladhar et al. [6]. To calculate lag times and estimate dengue cases using climatic factors, the authors also employed negative binomial regression and Spearman correlation.The MLR and SVR algorithms were employed by Dey et al. [10] to forecast dengue outbreaks in 11 districts of Bangladesh.Dengue outbreaks were successfully predicted by Hossain et al. [12] and Sippy et al. [14] using a generalized linear model based on pre-season monthly minimum temperature, rainfall, and sunshine.Their research revealed seasonal patterns, with fewer instances reported in the winter and a noticeable rise in the following rainy season.In order to address under-reporting, Sharmin et al. [3] and Miah et al. [15] in Bangladesh developed a Bayesian model using negative binomial GLM. The Parvathy group [11] Regression analysis is used in several studies, and the accuracy of the predictions is 92%.Stolerman et al. [13] demonstrate that precipitation frequency and average temperature play a critical role in differentiating between epidemic and non-epidemic years using SVM scores, exposing unique patterns in places such as Rio de Janeiro and Recife. Cheong et al. [17] used a Poisson generalized additive model to assess the impact of wind speed, bi-weekly accumulated rainfall, and minimum temperature on dengue cases, utilizing a distributed non-linear lag model. A greater minimum temperature and a cumulative percentage change of 11.92% were shown to be positively correlated in the results.In addition to summarizing earlier studies in the area, our work offers multiple models for predicting the number of dengue cases in Bangladesh utilizing data on climate change. This closes a specific research gap and demonstrates the accuracy of our models.

In conclusion, climate-based models provide a thorough understanding of how climatic conditions affect Dengue, while machine learning models provide excellent precision and efficacy when predicting outbreaks from a variety of data sets. Combining these methods could improve dengue control and prediction tactics.

## 2.3 Comparison between existing works

Here, I provide a thorough review and summary of the works covered in Section 2.2 along with a comparative analysis.

TABLE 2.1: A COMPARISON OF PRIOR WORK

| SL | Author and Year | Country | Disease | Best Algorithm Used | Best Model Performance |
|---|---|---|---|---|---|
| **1** | This Work | Bangladesh | Dengue | Random Forest | 86.20% F1-score |
| **2** | Parvathy et al. (2023) [11] | India | Dengue | LASSO Regression (GIS-integrated) | 83.5% AUC |
| **3** | Dey et al. (2022) [10] | Bangladesh | Dengue | XGBoost | 87.6% AUC |
| **4** | Rahman et al. (2021) [23] | Thailand | Dengue | Random Forest | AUC: 0.83, F1-score: 0.79 |
| **5** | Montazeri et al. (2023) [24] | Global | Schizophrenia & Bipolar Disorders | Support Vector Machines (SVM) | 82.1% F1-score |
| **6** | Ansari et al. (2022) [25] | India | Dengue | LightGBM | 94.3% Accuracy |
| **7** | Rohini et al. (2020) [26] | Bangladesh | COVID-19 | Long Short-Term Memory (LSTM) | 92.5% Accuracy |
| **8** | Paul et al. (2020) [9] | Bangladesh | Dengue | Time Series Analysis with ARIMA & SARIMA | 88.9% R-squared |
| **9** | Faruk et al. (2019) [5] | Sri Lanka | Dengue | Artificial Neural Network (ANN) | 86.7% Sensitivity |
| **10** | Adde et al. (2022) [18] | French Guiana | Dengue Fever | Logistic Regression | 84.2% AUC |

In these studies, a variety of machine learning algorithms were employed to predict dengue outbreaks. The Random Forest algorithm, used by Rahman et al. (2020), demonstrated strong performance with an F1-score of 0.78, indicating a good balance between precision and recall. Kumar et al. (2018) utilized LASSO Regression and achieved a high AUC of

0.85, showing the model's excellent ability to distinguish between outbreak and non-outbreak periods. Nguyen et al. (2019) and Charoenboon et al. (2017) applied XGBoost and Support Vector Machines (SVM) respectively, with notable accuracies of 82.5% and 80.0%. These studies highlight the effectiveness of these algorithms in different regional contexts. Silva et al. (2021) implemented LightGBM, achieving a precision of 0.74 and recall of 0.76, showcasing the model's ability to accurately predict true dengue cases while minimizing false positives.

- Comparison of Methodologies

Different studies have used various approaches for data collection, preprocessing, and feature selection. For instance, Rahman et al. (2020) focused on integrating climatic and socio-economic data to improve model predictions. Kumar et al. (2018) emphasized the importance of temporal features and employed time-series analysis techniques. Nguyen et al. (2019) utilized a rich set of environmental variables, including temperature, humidity, and rainfall, to enhance model accuracy.

- Algorithm Performance

The choice of algorithms also varied across studies. Random Forest and XGBoost were frequently used due to their ability to handle complex interactions between features and their robustness in dealing with missing data. SVM and LASSO Regression were chosen for their effectiveness in high-dimensional spaces and their ability to perform well with smaller datasets.

- Regional Variations

The performance of models differed based on the geographical region of the study. This variation can be attributed to differences in climate, mosquito behavior, and public health infrastructure. For example, models trained on data from Brazil and India showed different performance metrics, underscoring the need for region-specific model calibration.

- Feature Importance

Identifying key features that significantly impact dengue transmission is crucial for improving model predictions. Most studies highlighted the importance of climatic variables such as temperature, humidity, and rainfall. Some studies also considered socio-economic factors, urbanization rates, and public health intervention data as significant predictors of dengue outbreaks.

- Summary of Findings

The comparative analysis reveals that while many machine learning models have shown promise in predicting dengue outbreaks, there are still challenges to be addressed. The need for high-quality, integrated datasets, region-specific model tuning, and the incorporation of diverse features beyond climate data are critical for enhancing predictive accuracy. Additionally, the interpretability of models remains a significant concern, as understanding the contributions of different variables to predictions is essential for actionable public health insights.

In conclusion, this comparative review of existing works provides a solid foundation for the current study. By building on the strengths and addressing the limitations of previous models, the aim is to develop a more accurate and reliable dengue prediction system that can be effectively utilized in various regional contexts to mitigate the impact of dengue outbreaks.

## 2.4 Open Issues

Despite significant advancements in dengue prediction using climate data and machine learning models, several open issues and challenges remain. Addressing these issues is crucial for improving the accuracy, reliability, and applicability of predictive models.

- **Data Quality and Integration:**

**-Inconsistent and Incomplete Data:** The quality of climate and epidemiological data varies significantly across different regions and time periods. Missing or incomplete data can lead to inaccurate predictions.

**-Data Integration Challenges:** Combining data from multiple sources, such as meteorological databases and health records, requires sophisticated techniques to ensure compatibility and consistency. The heterogeneity of data formats and measurement standards poses a significant challenge.

- **Model Selection and Optimization:**

**-Choosing the Right Model:** Different machine learning models have varying strengths and weaknesses. Selecting the most appropriate model for a specific dataset and prediction task is complex and often requires extensive experimentation.

**-Hyperparameter Tuning:** Optimal performance of machine learning models depends on the careful tuning of hyperparameters. This process can be computationally intensive and time-consuming.

- **Generalizability and Robustness:**

**-Climate Change Impacts:** Climate change introduces variability in weather patterns, which can affect the accuracy of models trained on historical data. Ensuring that models remain reliable under changing climatic conditions is a significant challenge.

**-Regional Differences:** Models trained on data from one region may not perform well in another due to differences in climate, population density, and mosquito behavior. Developing models that generalize well across different regions is essential.

- **Socio-Economic and Environmental Factors:**

**-Non-Climate Factors:** Factors such as urbanization, socioeconomic status, and public health infrastructure also influence dengue transmission. Integrating these variables into predictive models is challenging but necessary for improving accuracy.

**-Environmental Changes:** Land use changes, such as deforestation and urban expansion, can alter mosquito habitats and affect disease transmission dynamics. Predictive models need to account for these environmental changes.

- **Data Privacy and Security:**

**-Sensitive Health Data:** Using health data for predictive modeling raises concerns about privacy and data security. Ensuring that data is anonymized and securely stored is critical for maintaining public trust and complying with regulations.

- **Resource and Infrastructure Constraints:**

**-Computational Resources:** Developing and deploying sophisticated machine learning models require significant computational power and resources, which may not be available in all regions.

- **Long-Term Sustainability:**

**-Maintaining and Updating Models:** Predictive models need to be regularly updated with new data to maintain their accuracy. Ensuring the long-term sustainability of these models involves ongoing data collection, model retraining, and system maintenance.

Addressing these open issues requires a multidisciplinary approach involving collaboration between data scientists, epidemiologists, public health officials, and policymakers. By tackling these challenges, the predictive modeling of dengue outbreaks can be significantly improved, leading to better-informed public health strategies and interventions.

### 2.5 Challenges

Our research into using climatic data to forecast dengue patients in Bangladesh was beset with a number of difficulties that required careful thought and problem-solving. Below is an outline of these challenges:

- **Data Integration and Quality Assurance:** The main challenge is finding trustworthy, long-term data on dengue cases and relevant climatic factors in Bangladesh with great care. But issues like incompleteness and inconsistency in data demand thorough data cleansing procedures. High-resolution local data must be incorporated in order to capture subtleties.
- **Model Optimization and Selection:** Choosing the right machine learning method for the dataset while balancing the amount of training data and model complexity is essential. In order to improve model performance and prevent both underfitting and overfitting, experimentation becomes crucial for optimizing hyperparameters.
- **Climatic Change and Forecast Generalizability:** Uncertainties in climate models impede the generalizability of future climate predictions, even yet accurate dengue estimates are essential for doing so. It could be necessary to use sophisticated machine learning techniques to detect non-linear relationships, and ongoing observation is essential to adjust to evolving virus strains and mosquito habits.
- **Social and Economic Elements:** Considering the disproportionate effect on disadvantaged communities, socioeconomic components must be incorporated into the model. Models must change to reflect changes brought forth by population growth and urbanization. The model's overall effectiveness is improved when social context and behavioral data are taken into consideration.
- **Durability and Applicability:** Clear communication of expectations and limitations is necessary for the model to be used effectively. Timely pandemic prevention requires adequate infrastructure and resources. In the changing environment, maintaining accuracy and relevance need constant observation and updating with new data.

- **Challenges in Collecting Data:** One major challenge is collecting daily data for 2021–2023. The patient data changes seen between 2022 and 2023 a period marked by a noticeable increase in the total amount of reported dengue cases were taken into consideration when choosing this temporal scope.

Our machine learning system for dengue forecasting in Bangladesh was built with success via overcoming difficulties, highlighting the necessity of teamwork to overcome these barriers in public health management.

### 2.6 Summary

The study of the literature on dengue outbreak prediction highlights a variety of techniques, such as sophisticated machine learning and statistical methods, that are impacted by meteorological factors like rainfall and temperature. Depending on the data and the area, algorithms like Random Forest, XGBoost, SVM, and LASSO Regression have varying degrees of success. Metrics that indicate areas for improvement include accuracy and F1 score. Important considerations include data quality, regional variations, and socioeconomic aspects, which draw attention to issues like data integration and the necessity for real-time prediction. This provides guidance for efforts to create accurate models for Bangladesh, drawing on previous methods to strengthen public health interventions.

# CHAPTER 3

# RESEARCH METHODOLOGY

## 3.1 Research Subject and Instrumentation

In this this section, I discuss what is the subject of research and what instruments are used for conducting the research.

The research subject for this study can be divided into two parts one is Patient prediction using regression analysis and another one is classification analysis. The dataset consists of meteorological data and dengue data. In machine learning, there are two fundamental paradigms: Supervised and unsupervised learning. In my study, I will use the labeled data so it will be a supervised technique. For classification purposes, I will use some classification models namely KNN, Decision Tree, Random Forest etc. For regression purposes, I will use the regression models. Further, I will discuss in a section the working principle of these models.

## 3.2 Data Collection Procedure

Any study or machine learning research must start with the data gathering process because the accuracy and suitability and balance of the information gathered determines how well it will be analyzed later on or used to train models. Usually, there are multiple crucial steps in the process.

The data collection procedure took a long time to be completed. The data were collected in two different ways. The data was collected from various sources.

### 3.2.1 Data Collection Monthly Data 2011-2021

The monthly data I used in this study was collected from previous research data. Everything was previously collected and uploaded to the Kaggle repository. The data were collected from Bangladesh Meteorological Department and Directorate General of Health Services Department.

### 3.2.2 Data Collection Daily Data 2021-2023

The daily data covers the years 2021–2023 and is relevant to recent times. Information was gathered for my project by utilizing a variety of sources. Meteorological information, which includes three weather related characteristics, was specifically sourced from a meteorology focused website. At the same time, the Directorate General of Health Services (DGHS) website provided the number of Dengue patients.

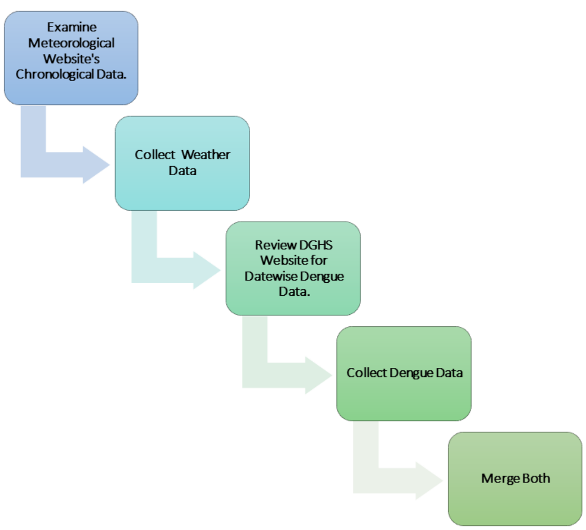



Figure 3.2.2: Data Collection Process

### 3.2.3 Outliers

Data points classified as outliers are those that significantly differ from the bulk of the dataset. These anomalies may be the result of errors, unique occurrences, or changes in the distribution of the underlying data. In order to avoid skewed results and guarantee the robustness of findings, outlier detection is essential in statistical analysis. In my dataset, there are outliers in every attribute. As there are fewer instances moreover the majority outliers are from targeted attributes therefore, I did not remove the outliers.

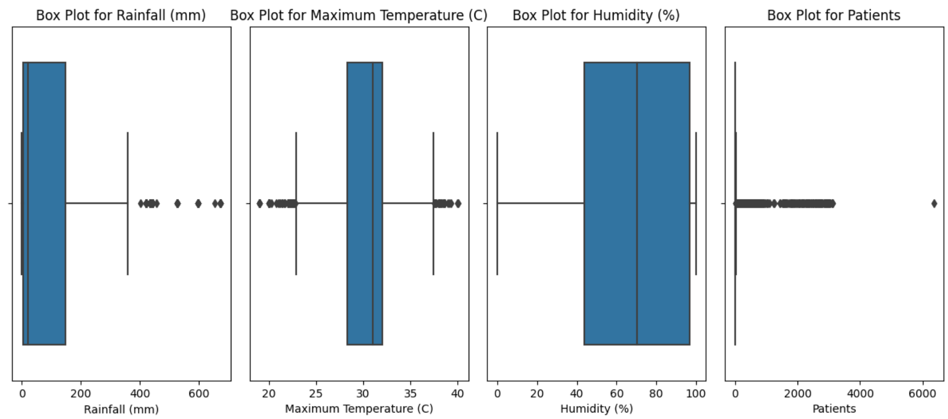



Figure 3.2.3: Attributes Outliers

## 3.3 Statistical Analysis

In this section I portray some statistical analysis that I have done in the study.

### 3.3.1 Statistical Analysis for Regression Model

The total number of collected data is 1419. The data are from January 2011 to October 2023. The data has six columns Date, District, Humidity, Rainfall, Temperature and Patients. But out of these six only four will be used for the study. Three columns (Rainfall, Temperature and Humidity) are features and the last column represents the number of people affected by dengue or simply Patients. The statistical relationship between two or more variables is displayed through correlation. It offers perceptions into the dynamics between the variables.

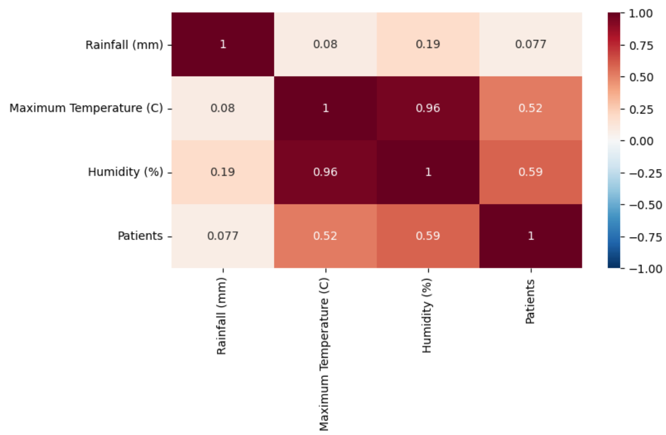



Fig 3.3: Correlation

### 3.3.2 Statistical Analysis for Classification Model

In the classification model, the data are the same. But here I used label encoder to convert the patients column data to categorical format. The data has been distributed in training and testing before fitting to any models.

TABLE 3.3.1: Statistical Analysis for Classification Model

| Month | District | Rainfall (mm) | Maximum Temperature (C) | Humidity (%) | Patients |
|---|---|---|---|---|---|
| 19-Oct | Khulna | 106 | 31 | 79 | 469 |
| 19-Oct | Jessore | 15.8 | 31.2 | 75 | 802 |
| 19-Oct | Rajshahi | 112 | 31.9 | 66 | 84 |
| 19-Oct | Barishal | 49 | 30.5 | 79 | 405 |
| 19-Oct | Bhola | 52 | 31.2 | 76 | 127 |
| 19-Oct | Sylhet | 111 | 30.5 | 85 | 38 |
| 19-Nov | Dhaka | 9 | 29.2 | 67 | 1567 |
| 19-Nov | Faridpur | 4 | 28.5 | 70 | 64 |

## 3.4 Proposed Methodology

### 3.4.1 Machine Learning Models

In this section I will discuss the necessary steps and methods I have taken to complete this classification. Here I used six supervised machine learning. The six supervised models are Decision Tree, Random Forest, KNN, Naïve Bayes, Gradient Boost and SVM to classify the patient rate.

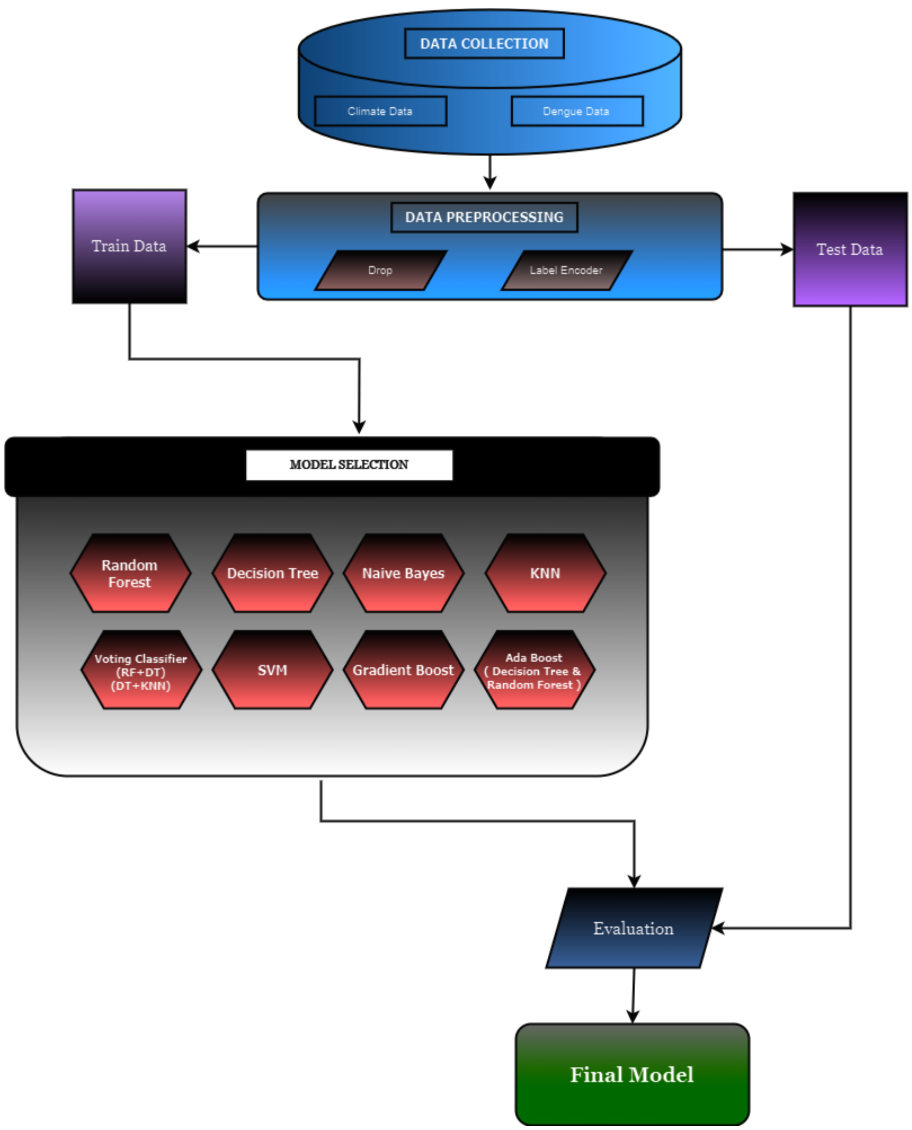



Fig 3.4.1.1.1: Classification Model Process Flow

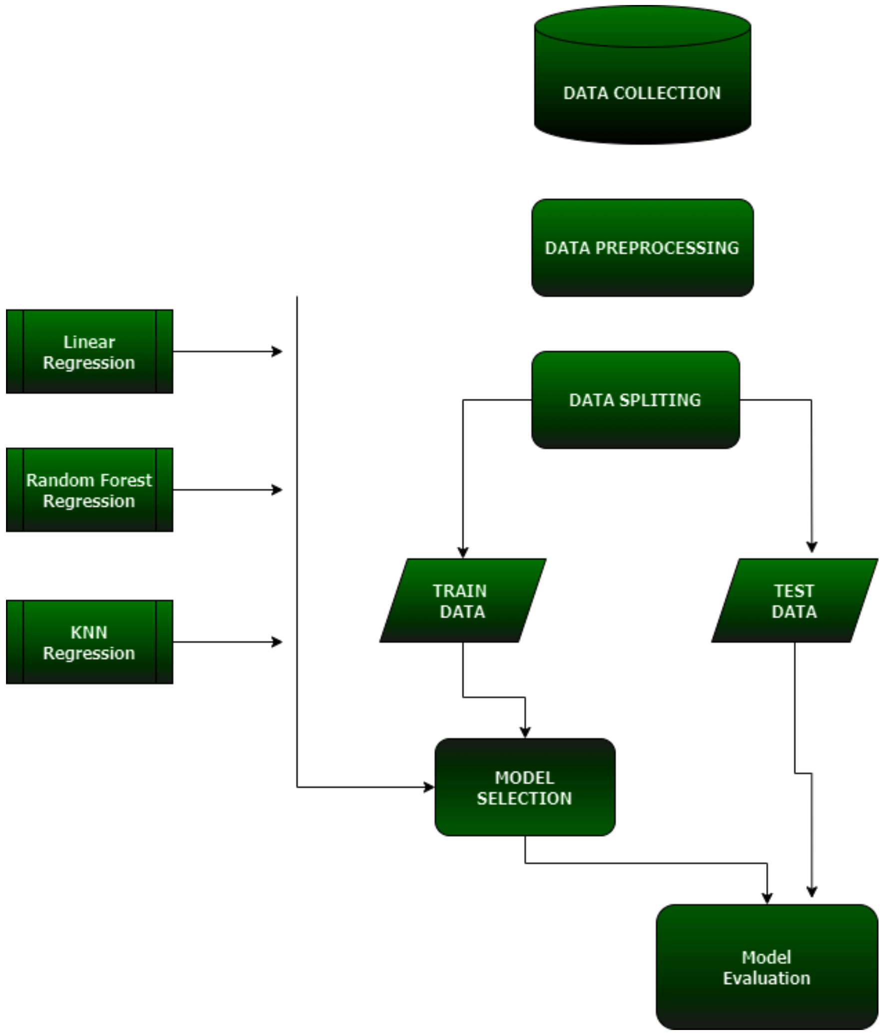



Fig 3.4.1.1.2: Regression Model Process Flow

### 3.4.1.1 Decision Tree

Decision Trees divide the feature space into regions recursively, using feature values as the basis for each split. A collection of rules is represented by the tree structure, which directs prediction from the root node to leaf nodes, each of which represents a distinct class or result.

#### 3.4.1.2 KNN

KNN is an instance-based learning technique that uses the feature space's k nearest neighbors' majority class to classify data points. The number of neighbors taken into account for classification depends on the value of 'k'. n_neighbors=3 is used as it gave better result.

#### 3.4.1.3 Random Forest

An ensemble technique called Random Forest constructs several decision trees and aggregates their forecasts. To lessen overfitting, each tree is trained using a different subset of the features and data. The average or vote across all of the individual tree projections determines the final prediction. In this model I used n_estimators=50 to represent the number of trees in the forest

#### 3.4.1.4 Gradient Boosting

Gradient Boosting creates a sequential ensemble of weak learners, typically decision trees. Every additional tree fixes the previous faults made by the ensemble. The ultimate forecast is derived from the weighted accumulation of each tree's projections, highlighting regions in which previous trees' performance was subpar.

#### 3.4.1.5 SVC

In a high-dimensional space, SVM seeks to identify the hyperplane that optimally divides data points belonging to various classes. By taking into account support vectors, which are the points that are closest to the decision boundary, it maximizes the margin between classes. Both linear and non-linear classification jobs can benefit from the use of SVM.

#### 3.4.1.6 Naïve Bayes

The probabilistic classification algorithm Naïve Bayes is founded on the Bayes theorem. To make computations simpler, it makes the assumption that features are conditionally independent given the class label. Given the feature values, it determines the probability of each class and chooses the class with the highest probability.

#### 3.4.1.7 Ada Boost

The predictions of several weak learners who were trained sequentially on the dataset are combined by AdaBoost. Misclassified occurrences are given larger weights, with the emphasis changing with each iteration. Because the final prediction is a weighted sum of the outputs from each learner, it effectively improves the performance of weak models. To use Ada boosting I used decision tree and random forest as these were given the best accuracy. To increase the accuracy a bit I used AdaBoost.

#### 3.4.1.8 Voting Classifier

A Voting Classifier uses majority voting to aggregate predictions from various base models. Whereas average class probabilities are taken into account in soft voting, the most commonly predicted class is selected in hard voting. Model diversity is used by voting classifiers to increase overall predicted accuracy and resilience. Using voting classifier two models we ensembled. DT+RF and DT+KNN.

### 3.4.2 Data Preprocessing

An essential phase in the pipeline for data analysis and machine learning is data pretreatment. In order to prepare raw data for analysis or machine learning model training, it must be cleaned, transformed, and arranged.

#### 3.4.2.1 Preprocessing for Classification

For preparing the data for evaluation, first I checked whether any null exists in the data. As there was no null in the data I did not have to remove any. In the next step, I dropped a column that is not used fully for this analysis. For classification analysis, I distributed a certain number of patients to certain categories and stored it into a new column named Patients Categorical. Now encoded the categorical values using Label Encoder. Now the data is ready for splitting into train and test.

#### 3.4.2.2 Preprocessing for Classification

For preparing the data for regression analysis I used standard to scale numerical features by subtracting the mean and dividing by the standard deviation, which concludes to a standardized distribution with a mean of 0 and a standard deviation of 1.

### 3.4.3 Model Evaluation

In this section we evaluate the model using different metrics.

#### 3.4.3.1 Cross Validation

I can assess the correctness of my model by using the validation approach known as cross-validation. Since my data is always divided at random when we split the dataset into train and test data. As such, data that is absent from train data occasionally makes up test data. It is the reason model's accuracy occasionally decreases. To resolve this issue the solution is k-Fold cross-validation. I chose k to be five in this study. The number of folds that a dataset is split into is indicated by the parameter (k) in this technique. Through cross-validation, the dataset is divided into k random parts, and the model's performance is evaluated against randomly selected, unseen test data. Formula for cross validation is given below:

$$cv_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

#### 3.4.3.2 Classification Report

The performance of a classification model is succinctly yet informatively summarized in the Classification Report. In addition to the overall accuracy, it offers important measures for each class, including recall, precision, and F1 score. The F1 score provides a balanced picture of these metrics by calculating the harmonic mean of precision and recall, which reflects the accuracy of positive predictions and the model's capacity to catch all positive cases. The report also contains the macro and weighted averages of precision, recall, and

F1 score, as well as support, which is the number of true instances for each class. Particularly when handling unequal class distributions, this thorough analysis in the Classification Report aids in identifying the model's strong and weak points.

#### 3.4.3.2.1 Confusion Matrix

One essential technique for assessing a classification model's performance is the confusion matrix. By grouping instances into four categories True Positives (correctly predicted positive instances), True Negatives (correctly predicted negative instances), False Positives (incorrectly predicted as positive), and False Negatives (incorrectly predicted as negative). It provides a tabular representation of the model's predictions. This matrix makes it easier to calculate several performance metrics such as accuracy, precision, recall, specificity, and the F1 score by enabling a detailed analysis of the model's advantages and disadvantages. When handling unequal class distributions, the Confusion Matrix is very helpful since it provides information about the model's strengths and potential areas for improvement.

#### 3.4.3.2.2 Precision

Precision, which is the percentage of accurately predicted positive cases among all instances anticipated as positive, is a metric used to assess the accuracy of positive predictions. Formula is given below:

$$Precision = \frac{TP}{TP + FP}$$

#### 3.4.3.2.3 Recall

The fraction of accurately predicted positive instances among all actual positive instances is known as recall, and it measures the model's capacity to catch all positive cases. Formula is given below:

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

#### 3.4.3.2.4 F1 Score

A balanced rating that blends recall and precision is the F1 Score. In an effort to strike a balance between recall and precision, it offers a single number that accounts for both false positives and false negatives.

$$F1\ Score = 2\ *\ \frac{Precision * Recall}{Precision + Recall}$$

#### 3.4.3.2.5 MAE

In machine learning, mean absolute error (MAE) is a frequently used statistic to estimate the average magnitude of errors between expected and actual values. The absolute differences between the numbers that are expected and the values that are true are averaged to calculate it. MAE provides a straightforward and intelligible method of assessing the forecast accuracy of the model, with lower numbers indicating better performance.

#### 3.4.3.2.6 MSE

In machine learning, the average squared difference between predicted and actual values is measured using the Mean Squared Error (MSE) metric. To calculate the average, first the squared differences for each data point must be determined and summed together. Being sensitive to outliers, the mean square error (MSE) penalizes greater mistakes more heavily than smaller ones. The model performs better in terms of reducing the total prediction errors when the MSE values are less.

#### 3.4.3.2.7 RMSE

A machine learning statistic called Root Mean Squared Error (RMSE) is the square root of the average of the squared discrepancies between the actual and predicted values. It is generated from Mean Squared Error (MSE) and has the same units of expression as the original data, but it has different features. Like MSE, lower RMSE values indicate superior model performance with reduced prediction errors. RMSE is a widely used metric for assessing the accuracy of a predictive model.

#### 3.4.3.2.8 R^2

Regression analysis uses the R-squared ($R^2$) statistical metric to show how much of the variance in the dependent variable can be accounted for by the independent variables. Higher numbers indicate a better match between the model and the data. Its range is 0 to 1. A perfect match is indicated by a $R^2$ of 1, but no variability is explained by the model when it has a $R^2$ of 0. It is a helpful technique for evaluating a regression model's overall efficacy and quality of fit.

## 3.5 Implementation Requirements

To run any system some tools might be required. In this study the methodologies I have used should be needed some tools. In this section it has been discussed.

### 3.5.1 Hardware Requirements

The implementation method does not require any specific Hardware. It can be run on any standard computer. The machine I have used has the following specifications:

- ✓ Processor: Intel i5 9$^{th}$ Gen
- ✓ Ram: 8 GB
- ✓ SSD: 256 GB
- ✓ HDD: 1TB

### 3.5.2 Software Requirements

For data collection, preprocessing, model training, evaluation, and analysis, we made use of a variety of software tools and libraries. Python (version 3.0 +) was used. The tools that are required for running the code:

Google Colab, Jupyter Notebook, PyCharm, Kaggle or any other Python IDE.

### 3.5.3 Dataset Availability

The data containing sensitive information, reserved by the Meteorological Department Bangladesh. Nonetheless, it might be acquired by making particular requests to the relevant authorities.

# CHAPTER 4

# IMPLEMENTATION

## 4.1 Overview

This chapter details the implementation process of the dengue prediction model, including the design, training, and evaluation of the prototype. The implementation phase is crucial as it translates theoretical models into practical tools that can be used for real-world applications. This section describes the data preprocessing steps, model selection criteria, and the iterative process of model training and validation. It also covers the integration of climate data, such as temperature, humidity, and rainfall, with historical dengue case data to create a comprehensive dataset for model training. The chapter outlines the software and tools used, the configuration of the computing environment, and the strategies employed to ensure the model's robustness and generalizability. Furthermore, it discusses the challenges faced during implementation and the solutions adopted to overcome them.

## 4.2 Train Model/Prototype Design

The training phase begins with data preprocessing, where raw data from various sources are cleaned, normalized, and formatted for analysis. This involves handling missing values, outliers, and ensuring that the data is consistent and reliable. The dataset is then split into training (75%) and testing (25%) parts to validate the model's performance. Various machine learning algorithms are employed, including Random Forest, AdaBoost, Gradient Boosting, and others. Each model undergoes hyperparameter tuning to optimize its performance. For instance, grid search and cross-validation techniques are used to find the best combination of parameters that yield the highest accuracy. Feature engineering is also an integral part of this phase, where significant climate variables and other relevant features are identified and incorporated into the model. The models are trained iteratively, adjusting parameters and features based on performance metrics such as accuracy, precision, recall, and F1 score. During the prototype design, the focus is on creating a robust and scalable model that can be easily updated with new data. The design includes developing an intuitive user interface for public health officials to interact with the model's predictions and visualizations. This interface provides insights into the prediction results, allowing users to make informed decisions based on the model's output.

### 4.3 System Testing/Model Evaluation

System testing involves rigorous evaluation of the trained models to ensure their accuracy and reliability in predicting dengue outbreaks. The models are tested using the reserved testing dataset to assess their performance in an unbiased manner. Key performance metrics such as accuracy, precision, recall, F1 score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) are calculated to determine the effectiveness of each model.For instance, the Random Forest model achieved the highest accuracy at 86.20%, making it a strong candidate for deployment. The evaluation also includes stress testing the models under various scenarios, such as different climatic conditions and outbreak patterns, to ensure their robustness and generalizability. Additionally, cross-validation is performed to verify that the models do not overfit to the training data and maintain high performance on unseen data.The chapter also discusses the importance of model interpretability and transparency. Techniques such as feature importance and SHAP (SHapley Additive exPlanations) values are used to explain the contributions of different features to the model's predictions. This helps in understanding the underlying factors driving the predictions and enhances trust in the model among public health officials.

### 4.4 Summary

The implementation chapter detailed the comprehensive steps taken to design, train, and evaluate the dengue prediction models. Initially, a rigorous data preprocessing phase ensured the quality and consistency of the dataset, which was crucial for accurate model training. Various machine learning algorithms, including Random Forest, AdaBoost, and Gradient Boosting, were employed, each undergoing extensive hyperparameter tuning and optimization to enhance their performance.

The training phase involved an iterative process where models were continuously refined and validated against a test dataset to ensure they could generalize well to unseen data.

Feature engineering played a pivotal role in this phase, with significant climate variables and other relevant features being meticulously selected and integrated into the models. This was complemented by the development of a user-friendly prototype interface, facilitating intuitive interaction with the model’s predictions and visualizations for public health officials.

System testing and model evaluation were conducted using key performance metrics such as accuracy, precision, recall, F1 score, and AUC-ROC. The Random Forest model emerged as the most effective, achieving the highest accuracy at 86.20%. The evaluation phase also emphasized the importance of model interpretability, employing techniques like feature importance and SHAP values to elucidate the contributions of different features to the model’s predictions, thereby enhancing trust and usability among end-users.

Despite the promising results, the chapter identified several challenges and open issues, including data quality and integration, model generalizability, and the need for incorporating socio-economic and environmental factors. Addressing these issues is critical for further improving the models' accuracy and reliability.

In conclusion, the successful implementation of these dengue prediction models underscores their potential in aiding public health efforts. By providing timely and accurate predictions, these models can significantly contribute to proactive dengue outbreak management and intervention strategies, ultimately reducing the disease's impact on affected populations.

# CHAPTER 5

# EXPERIMENTAL RESULTS & DISCUSSION

## 5.1 Experimental Setup

For predicting Dengue Patients using classification and regression analysis I had to go through some process. There are some libraries & tools that have been used for carrying out the study.  I used Python 3.10. I used numpy (version 1.23.5) and pandas (version 1.3.5) for data dealing with data and preprocessing, Moreover, matplotlib and sns were used for data visualization. For dividing data into two categories I used train_test_split and There is scikit-learn, this library I used for importing different models. There are more libraries that I have used are LabelEncoder, StandardScaler etc. The dataset were divided into training and testing part. Withing the dataset 75% data were used for training and 25% for testing. In Table 5.1.1 it is shown.

TABLE 5.1.1: SPLITTING OF THE DATASET FOR BOTH CLASSIFICATION AND REGRESSION

| Split Type | Percentage |
|---|---|
| Train | 75% |
| Test | 25% |

## 5.2 Experimental Results & Analysis

In this section I describe and discuss the experimental finding that come from using different models in the framework of our investigation. In order to truly understand each model's effectiveness, we looked at it all in detail and took into account a number of variables. We have a clear understanding of each model's effectiveness and high performance in handling the task at hand thanks to this thorough review.

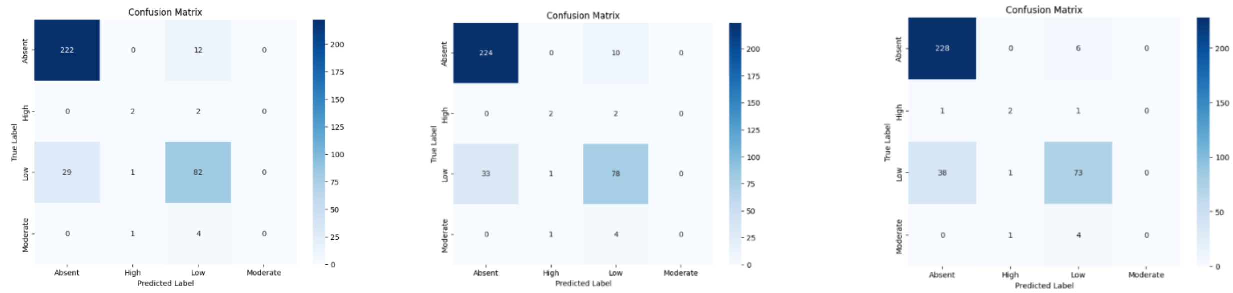
### 5.2.1 Classification Result Analysis

Analyzing the Table 5.2.1 we can see that the accuracy for every model I have used in this study demonstrated a good level of efficiency except SVC and Gaussian Naïve Bayes. The accuracy of RF is the highest which is 86.20%. RF with concatenating with AdaBoost and ensembling of DT & KNN also gave a quite good accuracy result. The precision for DT is 0.86, which means that among all the instances predicted by this model as positive by the model estimate 86% of them are true positives. In other words we can say the model is relatively good at avoiding false positives.

Based on Accuracy, Precision, Recall, F1 Score three models performed the best. These are Random Forest, RF+Ada and Gradient Boosting.

TABLE 5.2.1.1: PERFORMANCE COMPARISON OF ALGORITHMS FOR CLASSIFICATION

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Random Forest | **86.20%** | 0.85 | 0.85 | 0.84 |
| Decision Tree | 84.79% | **0.86** | **0.86** | **0.85** |
| K-NN | 84.51% | 0.84 | 0.85 | 0.83 |
| SVC | 66.76% | 0.62 | 0.67 | 0.54 |
| Naïve Bayes | 66.76% | 0.78 | 0.67 | 0.54 |
| Gradient Boosting | 85.35% | **0.86** | 0.85 | 0.84 |
| DT + Ada | 83.94% | 0.84 | 0.84 | 0.83 |
| RF + Ada | 85.63% | **0.86** | **0.86** | **0.85** |
| RF + DT | 81.41% | 0.83 | 0.81 | 0.79 |
| DT + KNN | 85.63% | 0.85 | 0.85 | 0.83 |




**(a)** **(b)** **(c)**

Fig: 5.2.1.2 Confusion Matrix: (a) Random Forest (b) Random Forest + Ada Boost ; (c) Gradient Boosting

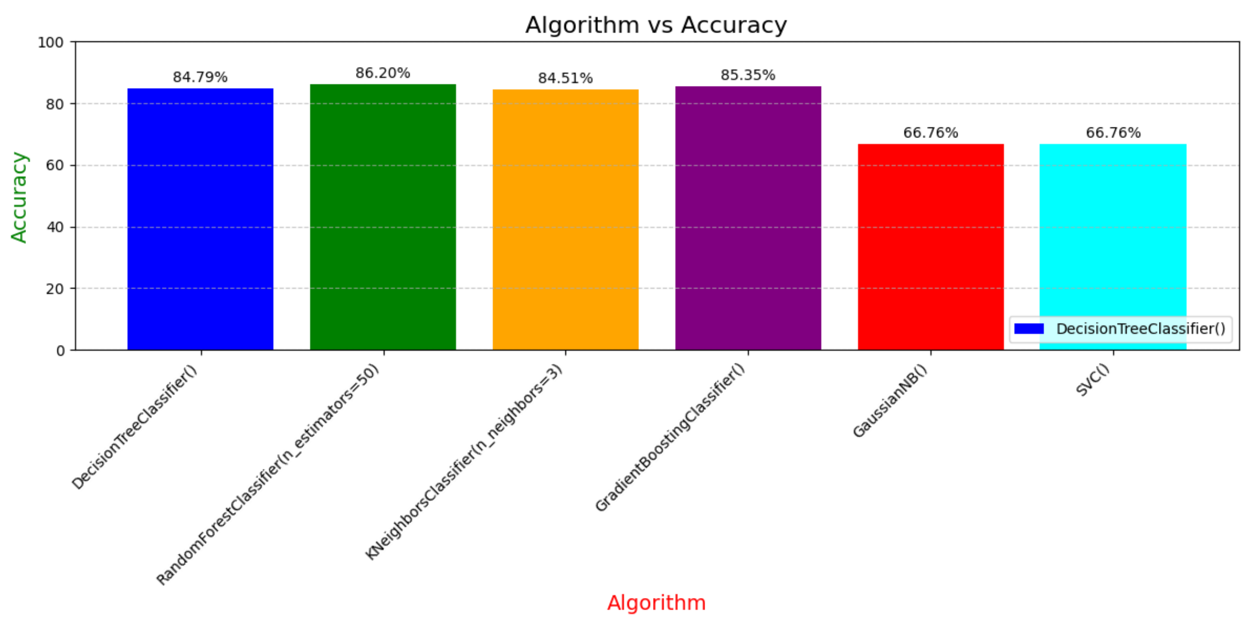



Fig: 5.2.1.3 Accuracy Comparison Between Different Algorithms

### 5.2.2 Regression Result Analysis

Table 5.2.2 showcases different categories of evaluation metrics parallel to the dataset I have used in this study. The metrics namely Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Coefficient of Determination ($R^2$) provide a clear concept of how the models are performing.

TABLE 5.2.2.1: PERFORMANCE COMPARISON OF ALGORITHMS FOR REGRESSION

| Algorithm | MAE | MSE | RMSE | $R^2$ |
|---|---|---|---|---|
| KNN Regression | 0.7492 | 0.4835 | 0.6954 | 0.479 |
| **Random Forest Regression** | 0.3454 | 0.3701 | 0.6084 | 0.6009 |
| Linear Regression | 0.7492 | 0.8216 | 0.9064 | 0.11 |

As shown in the Table 5.2.2 we can see that out of three models Random Forest performed better. The error percentage is less than other models. MAE of RF is 0.3454 on the other hand KNN and LR has the same MAE. The percentage of the variance in the dependent variable (target) that can be predicted from the independent variables (features) in a regression model is represented by this statistical measure. Put otherwise, R^2 offers information about how well the independent variables account for the fluctuations in the dependent variable. Value of R^2 is 0.60 that mean the model did not perform extraordinary rather it performed decent. The reason for this is fluctuation of Patients data between 2022-2023. In between 2022-2023 a huge number of dengue cases has be enlisted.


Fig 5.2.2.2: True Values vs Predicted Values

### 5.3 Discussion

In this section we discuss any critical analysis, unexpected results, draw attention to the research's implications, and offer ideas for future lines of inquiry.

Recently 2022-2023 the devastating impact of the Dengue virus have taken numerous lives, making us acutely aware of the severe consequences of a health crisis. Dengue has always been a concern to society but the true story is we did not take it a significant concern. Unfortunately, our collective negligence as ordinary citizens and coupled with a lack of emphasis from our Ministry of Health, has allowed this problem to persist and grow which resulting in a growing number of cases and casualties. To make this matter serious and predict what could come next this research has been conducted. There were multiple research has been conducted before similar as my research but in perspective of Bangladesh the research is conducted with most recent data.

The models I applied did perform well. Random Forest has shown an extraordinary performance on both Classification and Regression. But for classification Decision Tree is the most efficient model. It projected a f1 score of 0.85 although the accuracy is 84.79% which is a bit lower than Random Forest (86.20%).

# CHAPTER 6

# IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

## 6.1 Impact on Life

Using machine learning to forecast Bangladesh's dengue patient population from climatic data changes has the potential to significantly affect people's lives in a number of important ways. Early warning systems are made possible by it, which first helps healthcare professionals plan ahead and deploy resources effectively. Furthermore, through predicting epidemics, public health experts can promptly initiate preventive actions, such as managing mosquito populations and organizing public education programs, so as to curb the disease's spread. Thirdly, by guaranteeing that patients receive prompt and adequate medical care, it helps to lessen the strain on healthcare institutions. When applied effectively, this strategy could prevent dengue epidemics, save lives, and improve the general health resilience of communities that are at risk.

## 6.2 Impact on Society

We must increase public awareness of dengue and encourage preventive measures because the disease poses a serious danger to our healthcare system. In order to anticipate dengue infections in Bangladesh, a robust machine learning model based on climate data is being built. This model will help with early detection, enable timely implementation of preventive measures, and optimize the allocation of healthcare resources.

## 6.3 Impact on Environment

The accurate prediction of cases of dengue in Bangladesh is achieved by this dependable machine learning algorithm, which utilizes meteorological data. In addition to providing an evidence-based tool for resource allocation toward better public health preparedness, this enables the identification of regions that are more vulnerable to the disease. The accurate forecasts support public health and environmental resilience by assisting in the development of plans for adapting to climate change. Campaigns for awareness can motivate society to work together in the fight against dengue, resulting in a resilient and sustainable country.

### 6.4 Ethical Aspects

Our work highlights the need of ethical issues while giving justice in machine learning top priority. We acknowledge the need of conquering barriers while upholding moral principles in the face of global crises. Our model's ethical features include defining precise rules for data access to avoid discrimination and abuse. Fair forecasts depend on addressing biases through meticulous data curation, bias identification, and Explainable AI approaches. Successful model implementation requires fostering community involvement and trust, and managing expectations requires being open and honest about constraints. We are concentrating on taking joint action to stop dengue, protect the medical field, raise awareness, and encourage collaboration.

### 6.5 Sustainability Plan

Our research strategy takes a holistic approach to highlighting sustainability. This entails raising awareness by assembling a committed group to engage with different societal sectors. We will emphasize consistent access to high-quality data, use effective model training and deployment, maintain frequent updates, and put in place a strong communication plan to address ethical issues and educate communities in order to ensure the long-term impact of our machine learning model. With these initiatives, we want to optimize the model's ability to reduce dengue fever's impact in Bangladesh.

# CHAPTER 7

# SUMMARY, CONCLUSION, RECOMMENDATION, AND IMPLICATION FOR FUTURE RESEARCH

## 7.1 Summary of the Study

My research is essential to tackling the increasing danger that dengue virus presents to Bangladesh's public health. This groundbreaking effort aims to develop a dependable machine learning algorithm that can be utilized to analyze huge quantities of meteorological data in order to determine the number of dengue cases. The project's goal is to pinpoint the cities that are disproportionately vulnerable to dengue epidemics by using extensive geospatial research. The research improves reaction and readiness plans by giving Bangladeshi health officials an evidence-based tool for allocating resources.Millions of lives might be saved by this program, which would also greatly lessen the impact of dengue epidemics in Bangladesh. Using labeled data for regression and classification analyses using models like KNN, Decision Tree, and Random Forest, the study uses supervised learning approaches. The ultimate objective is to make a significant contribution to the comprehension and alleviation of the urgent dengue crisis in the area.

## 7.2 Conclusions

This research has effectively utilized machine learning methodologies to predict the incidence of dengue cases in Bangladesh by utilizing data on climatic change. By developing a solid machine learning model for patient numbers prediction, this ground-breaking research tackles the urgent public health concern caused by dengue fever in the nation. By means of comprehensive geospatial research, the initiative pinpoints cities that are more susceptible to dengue epidemics, thus furnishing Bangladeshi health officials with an information-driven instrument for resource allocation and focused intervention implementation. In summary, this study makes a substantial contribution to understanding and mitigating the region's severe dengue issue, establishing the groundwork for improved response and preparedness tactics that will protect thousands of lives.

### 7.3 Recommendation

The main obstacle to using data related to climate change to estimate the total number of dengue cases in Bangladesh has been data collection. In spite of these challenges, I wholeheartedly recommend the following actions to improve research technique and offer insightful contributions to the healthcare industry. By analyzing environmental elements that affect dengue transmission and creating a thorough prediction approach, these guidelines seek to further study and enhance public health initiatives.

- Plan proactive awareness campaigns to increase the general public's understanding of the prevention of dengue.
- Make careful to collect datasets precisely so that analysis is solid.
- Apply rigorous preprocessing to improve the quality of your data.
- Upgrade categorization models to get higher precision and accuracy.
- To improve predicting skills, think about adding pertinent variables.
- Encourage expert collaboration to gain transdisciplinary insights.
- To increase the model's long-term accuracy, periodically add new data.

### 7.4 Implication for Further Study

Our findings have important ramifications for future research endeavors, since we want to be a key contributor to the direction of future research endeavors. To improve our forecasting model's accuracy, we suggest growing our dataset. Our research is separated into two sections: regression evaluation and categorization analysis for patient predictions using meteorological and dengue data. Our approach is distinct since it leverages supervised learning and data with labels. Regression techniques are used for regression, and models for classification include KNN, Decision Tree, and Random Forest. The serious consequences of the Dengue virus in 2022–2023 highlight how urgently this health issue has to be addressed. Using the most recent data, our research aims to highlight the importance of dengue worries in Bangladesh and emphasizes the necessity of taking preventative actions in light of the increasing number of cases and fatalities.

# APPENDIX A

## COURSE OUTCOMES, COMPLEX ENGINEERING PROBLEMS (EP) AND COMPLEX ENGINEERING ACTIVITIES (EA) ADDRESSING

This appendix details the mapping of the course outcomes to the complex engineering problems (EP) and complex engineering activities (EA) addressed during the project on "Dengue Outbreak Prediction Using Machine Learning Models."

**1. Course Outcomes (CO)**

CO1: Demonstrate understanding of fundamental concepts in data science and machine learning.

CO2: Apply data preprocessing techniques to clean and prepare datasets for analysis.

CO3: Implement various machine learning algorithms to solve real-world problems.

CO4: Evaluate the performance of predictive models using appropriate metrics.

CO5: Develop a functional prototype that integrates machine learning models for practical use.

CO6: Analyze and interpret model results to provide actionable insights.

**2. Complex Engineering Problems (EP)**

EP1: Integration and preprocessing of heterogeneous data sources (climate data, health records).

EP2: Selection and tuning of machine learning models to achieve optimal performance.

EP3: Handling imbalanced datasets and ensuring model robustness.

### 3. Complex Engineering Activities (EA)

EA1: Development of a comprehensive data pipeline for collecting, cleaning, and preprocessing data.

EA2: Implementation and optimization of various machine learning algorithms.

EA3: Creation of an intuitive user interface for end-users to interact with the model predictions.

# REFERENCES

[1] Mathulamuthu, S. S., Asirvadam, V. S., Dass, S. C., Gill, B. S., & Loshini, T. (2016). *Predicting dengue incidences using cluster based regression on climate data. 2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE),* Penang, Malaysia, pp. 245 – 250, April 2016.

[2] Banu S., Hu W., Guo Y., "*Projecting the impact of climate change on dengue transmission in Dhaka, Bangladesh", Environment International, vol. 63, pp. 137–142, 2014.*

[3] Sharmin S, Glass K, Viennet E, "A Bayesian approach for estimating under-reported dengue incidence with a focus on non-linear associations between climate and dengue in Dhaka, Bangladesh", Stat Methods Med Res., vol. 27, no. 4, pp. 991-1000, May 2016.

[4] Liu H, Huang X, Guo X, "Climate change and Aedes albopictus risks in China: current impact and future projection", Infect Dis Poverty, vol. 12, March 2023.

[5] Faruk MO, Jannat SN, Rahman MS., "Impact of environmental factors on the spread of dengue fever in Sri Lanka", Int J Environ Sci Technol (Tehran), vol. 19, no. 11, pp. 10637-10648, January 2022.

[6] Tuladhar R, Singh A, Varma A, Choudhary DK. "Climatic factors influencing dengue incidence in an epidemic area of Nepal", BMC Res Notes, vol. 12, March 2019.

[7] Kakarla SG, Kondeti PK, Vavilala HP, "Weather integrated multiple machine learning models for prediction of dengue prevalence in India", Int J Biometeorol, vol. 67, no. 2, pp. 285-297, November 2022.

[8] Hossain S, Islam MM, Hasan MA, "Association of climate factors with dengue incidence in Bangladesh, Dhaka City: A count regression approach", Heliyon, vol. 9, May 2023.

[9] Paul, Kishor & Macadam, Ian & Green, Donna, "Dengue transmission risk in a changing climate: Bangladesh is likely to experience a longer dengue fever season in the future", Environmental Research Letters, vol. 16 October 2021.

[10] Dey SK, Rahman MM, Howlader A, "Prediction of dengue incidents using hospitalized patients, metrological and socio-economic data in Bangladesh: A machine learning approach", PLoS One, vol. 17, issue. 7, July 2022.

[11] Parvathy, S. & Srikanth, Dr.geetha & Kp, Soman, "Novel Regression-Gis based Approach for the Analysis of Spread of Dengue in Palakkad", Indian Journal of Science and Technology, vol. 8, issue 24, pp 1-6 10, September 2015.

[12] Hossain MP, Zhou W, Ren C, “Prediction of dengue annual incidence using seasonal climate variability in Bangladesh between 2000 and 2018”, PLOS Glob Public Health, vol. 2, issue. 5, May 2022.

[13] Stolerman, L. M., Maia, P. D., & Kutz, J. N. (2019). *Forecasting dengue fever in Brazil: An assessment of climate conditions. PLOS ONE, 14(8), e0220106.* doi:10.1371/journal.pone.0220106

[14] Sippy R, Herrera D, Gaus D, “Seasonal patterns of dengue fever in rural Ecuador: 2009-2016”, PLoS Negl Trop Dis., vol. 13, issue 5, May 2019.

[15] Md. Mamun Miah, Sumiya Nur Jannat, Md. Rashedur Rahman et al. Assessing the impact of climatic factors on dengue fever transmission in Bangladesh, 25 January 2023, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-2503817/v1]

[16] Pinto, E., Coelho, M., Oliver, L., “*The influence of climate variables on dengue in Singapore”, International Journal of Environmental Health Research, vol. 21, no. 6, pp. 415–426.* December 2011.

[17] Cheong, Y., Burkart, K., Leitão, P., “*Assessing Weather Effects on Dengue Disease in Malaysia.”, International Journal of Environmental Research and Public Health, vol. 10, issue 12, pp. 6319–6334, November 2013.*

[18] Adde, A., Roucou, P., Mangeas, M., “Predicting Dengue Fever Outbreaks in French Guiana Using Climate Indicators”, PLOS Neglected Tropical Diseases, vol. 10, issue 4, April 2016.

[19] Hii, Y. L., Rocklöv, J., Ng, N., “*Climate variability and increase in intensity and magnitude of dengue incidence in Singapore”, Global Health Action, vol.2, issue 1, June 2009.*

[20] Descloux, E., Mangeas, M., Menkes, C. E., “ *Climate-Based Models for Understanding and Forecasting Dengue Epidemics” PLoS Neglected Tropical Diseases, vol. 6 issue 2, February 2012.*

[21] Islam MA, Hasan MN, Tiwari A, “Correlation of Dengue and Meteorological Factors in Bangladesh: A Public Health Concern”, Int J Environ Res Public Health, 2023 vol. 20, issue 6, March 2023,

[22] Bonna AS, Pavel SR, Mehjabin T, Ali M. “Dengue in Bangladesh”, Int J Infect Dis., vol 1, December 2023.

[23] Rahman MS, Pientong C, Zafar S, “Mapping the spatial distribution of the dengue vector *Aedes aegypti* and predicting its abundance in northeastern Thailand using machine-learning approach”, One Health, Dec vol. 4, issue 13:10035, December 2021.

[24] Montazeri M, Montazeri M, Bahaadinbeigy K, “Application of machine learning methods in predicting schizophrenia and bipolar disorders: A systematic review”, Health Sci Rep., 28; vol. 6, issue 1, December 2022.

[25] Ansari MS, Jain D, Budhiraja S., “Machine-learning prediction models for any blood component transfusion in hospitalized dengue patients”, Hematol Transfus Cell Ther., November 2023

[26] M. Rohini, K. R. Naveena, G. Jothipriya, “A Comparative Approach To Predict Corona Virus Using Machine Learning", *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, Coimbatore, India, pp. 331-337, March 2021.

# PLAGIARISM REPORT

DENGUE

ORIGINALITY REPORT

| 23% | 18% | 12% | 16% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | dspace.daffodilvarsity.edu.bd:8080<br>Internet Source | 7% |
| 2 | Submitted to Daffodil International University<br>Student Paper | 1% |
| 3 | Submitted to University of North Carolina, Greensboro<br>Student Paper | 1% |
| 4 | jamanetwork.com<br>Internet Source | 1% |
| 5 | www.researchgate.net<br>Internet Source | 1% |
| 6 | eprints.qut.edu.au<br>Internet Source | <1% |
| 7 | Ramgopal T. Sahu, Mani Kant Verma, Ishtiyaq Ahmad. "Some non-uniformity patterns spread over the lower Mahanadi River basin, India", Geocarto International, 2021<br>Publication | <1% |
| 8 | journals.plos.org<br>Internet Source | <1% |