A Comparative Study of Machine Learning and Geospatial Techniques for Analyzing Dengue Diffusion Patterns and Identifying Hotspots in Bangladesh.

by

Taskia Siddika
20301417
Shuria Akter Ethuna
20301055
Nafisa Ahmed Progga
24341096
Niamotullah Ratul
19301151
Mirza Fahad Bin Kamal
20101399

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
January 31, 2025

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Niamotullah

Niamotullah Ratul

19301151

মির্জা ফাহাদ বিন কামাল

Mirza Fahad Bin Kamal

20101399

Taskia Siddika

Taskia Siddika

20301417

Nafisa Ahmed Progga

Nafisa Ahmed Progga

24341096

Shuria Akter Ethuna

Shuria Akter Ethuna

20301055

Approval

The thesis titled "A Comparative Study of Machine Learning and Geospatial Techniques for Analyzing Dengue Diffusion Patterns and Identifying Hotspots in Bangladesh." submitted by

1. Taskia Siddika (20301417)
2. Shuria Akter Ethuna (20301055)
3. Nafisa Ahmed Progga (24341096)
4. Niamotullah Ratul (19301151)
5. Mirza Fahad Bin Kamal (20101399)

Of Fall, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 31, 2025.

Examining Committee:

Supervisor:
(Member)

Md. Golam Rabiul Alam, PhD

Professor
Department of Computer Science and Engineering
School of Data and Science
BRAC University

Co-Supervisor:
(Member)

RRahman

Rafeed Rahman

Lecturer
Department of Computer Science and Engineering
School of Data and Science
BRAC University

Program Coordinator:
(Member)

Rabiul

Md. Golam Rabiul Alam, PhD

Professor
Department of Computer Science and Engineering
School of Data and Science
BRAC University

Head of Department:
(Chairperson)

SHKazi

Dr. Sadia Hamid Kazi

Chairperson And Associate Professor
Department of Computer Science and Engineering
School of Data and Science
BRAC University

Abstract

Dengue fever is still a major public health challenge in tropical and subtropical countries, especially in Bangladesh where epidemic has been a big threat to public health. In this paper, we have developed an integrative computational approach analyzing geographic information and employing data mining to forecast dengue spread and reveal vulnerable regions. Our reference methods include Ordinary Least Squares (OLS), Geographically Weighted Regression (GWR), Lasso Regression, Elastic Net Regression, Bidirectional Long Short Term Memory (BiLSTM), and DiffFlow as well as TabDDPM. The study builds on past results, climatic parameters, and population density to improve predictive performance. The data set used in this research was collected from the official web site of Directorate General of Health Services (DGHS) that made the data authentic. The proposed approach, as a result, provides significantly higher predictive performance than conventional statistical analysis based on the spatial and machine learning components. Furthermore, to categorize and prioritize the high-risk areas, we apply special methods of multiple criteria decision making – TOPSIS and VIKOR. We reveal that the proposed models based on machine learning methodologies are useful for identifying dengue fever hotspot areas and enlightening information for public health officials regarding timely application of control measures.. This study emphasises the necessity of the epidemiological and climate data coupled with the computational modeling to mitigate future outbreaks.

Keywords: Dengue, Diffusion Patterns, Machine Learning, Geospatial Analysis and Effective Distance Model.

Dedication

We would like to dedicate this research to our parents, to whom we are forever grateful.

Acknowledgment

Firstly, thanks to our beloved family members to whom we are, and always will be indebted. Secondly, to our supervisor Md. Golam Rabiul Alam for all his help and constant support of our endeavors. Thirdly, we would like to thank all the faculty members, staff, our peers, Research Assistant Mr. Nafiz Imtiaz Rafin, and our co-supervisor Mr. Rafeed Rahman, along with other stakeholders related to BRAC University, for providing us an environment in which we were able to develop ourselves, learn to our fullest extent, and conduct our research properly.

Finally, and most importantly, we are deeply grateful to Allah for granting us the strength, patience, and guidance to complete this work in time.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several abbreviation that will be later used within the body of the document

α	Alpha
β	Beta
γ	Gamma
∞	Infinity
θ	Theta
ANN	Artificial Neural network
CFG	Case Fatality Grade
DT	Decision Tree
GBR	Gradient Boosting Regression
GNB	Gaussian Neighbor Boundaries
KNN	K-Nearest Neighbor
LM	Levenberg Marquard
LR	Logistic Regression
LVQ	Learning Vector Quantizatio
MAE	Mean Absolute Error
MLR	Multiple Linear Regression
NN	Neural Networks
RF	Random Forest
$RMSE$	Root-Mean-Square-Error
ROC	Receiver Operating Characteristic
SCG	Scaled Conjugate Gradient
SVC	Support Vector Classifier

SVR Support Vector Regression

$XGBR$ Extreme and Light Gradient Boosting Regression

Chapter 1

Introduction

Dengue fever is an infectious tropical disease transmitted through the yellow fever mosquito and is one of the world's most important emerging diseases that has introduced a new era of significant threats in public health. The disease is transmitted through Aedes aegypti and Aedes albopictus mosquitoes and has attributed outbreak incidences that have increased to nationwide complications in the recent past years. Special focus on Bangladesh, the dengue fever has raised its morbidity and mortality rates remarkably in recent years. The DGHS has also recorded that the last few years have witnessed a record high in infections and hence the need for modelling to help to detect the likely outbreak. This superseding of nature through an epidemiological understanding of the future of dengue has been done using post hoc statistical regression analysis and/or case histories. However, the said approaches do not take into consideration such aspects as non-linear temporal dependence of environmental factors and the distribution of density population and diseases. The use of machine learning and geospatial approaches provides a fresh and efficient solution which helps to analyze the outbreaks more profound and adaptive way.

1.1 Motivation

Dengue outbreaks have become a recurring public health challenge in tropical countries like Bangladesh, particularly during the rainy season when Aedes mosquitoes breed in standing water. With no specific treatment available, prevention remains the most effective strategy for controlling the virus. Identifying high-risk areas in advance can help authorities implement targeted interventions to mitigate the impact of outbreaks. The increasing frequency and severity of dengue outbreaks demand the development of advanced computational models to identify potential dengue hotspots and aid in public health planning. The situation in Bangladesh has been exacerbated by rapid urbanization, climate change, and inadequate public health infrastructure, making it crucial to integrate machine learning and geospatial techniques for improved hotspot detection.

Existing models for dengue diffusion primarily rely on spatio-temporal analysis, which may overlook potential hotspots in regions with no prior outbreak history. This study seeks to address this limitation by incorporating climatic parameters, population density, and patient information from hospital reports of dengue cases to enhance the accuracy of hotspot identification. Furthermore, the study applies multiple criteria decision-making techniques, such as TOPSIS and VIKOR, to categorize and prioritize high-risk areas. By providing a decision-support tool for health authorities, this study aims to enhance outbreak prevention strategies and optimize resource allocation for controlling dengue in vulnerable regions.

Key Objectives

- Design a refined computational algorithm by which Dengue cases can be predicted in a population other than depending on the standard temporal-spatial distribution model.
- Combine geospatial and machine learning methods to increase hotspot identification accuracy.
- To improve the model, include population density, climatic characteristics, and patient information from hospital records.
- Classify and rank high-risk locations using the multiple criterion decision-making techniques (VIKOR and TOPSIS).
- Give public health officials a decision-support tool to help them allocate resources effectively and carry out targeted treatments in high-risk areas.

1.2 Problem Statement

Over the past four to five years, dengue has become a major epidemic in Bangladesh, claiming many lives. According to UNICEF, in 2019, 0.1 million people were infected, and the death toll reached 164. Continuing this trend, in 2023, nearly 0.075 million people were affected, with the number of deaths rising to 352, as reported by the World Health Organization (WHO). This situation is alarming for a country with limited healthcare infrastructure. Since there is no effective treatment for dengue, the focus remains on symptom management rather than a cure. This makes prevention the most crucial strategy for controlling the disease.

Extensive research has been conducted on disease detection and epidemic prediction, with many studies producing highly reliable models. However, there is still a lack of significant research on dengue diffusion in Bangladesh, despite similar studies being conducted in other countries for various epidemic diseases, including dengue. Therefore, this study aims to analyze dengue diffusion patterns and identify potential dengue hotspots in advance. By achieving this, authorities can take proactive measures to mitigate outbreaks, making it easier for a country like Bangladesh to effectively manage this public health crisis.

1.2.1 Research Gaps

Despite the extensive research on dengue outbreaks, there are significant gaps in the existing studies, particularly in the area of dengue hotspot identification. While some studies focus on predictive models for dengue spread, these models often fail to accurately identify high-risk regions in a timely manner. Additionally, many existing approaches are limited to traditional epidemiological methods or use only one or two variables, neglecting the potential of integrating multiple data sources and advanced computational techniques. This research seeks to address these gaps by developing a more comprehensive approach that combines machine learning, geospatial analysis, and decision-making techniques.

- Limited Focus on Hotspot Identification: While many studies on dengue have used machine learning and geographical approaches for prediction, there is a lack of research specifically addressing the identification of dengue hotspots. Existing studies typically focus on the prediction of outbreaks, but fail to provide detailed methods for pinpointing high-risk areas within a region.
- Insufficient Integration of Climatic and Epidemiological Data: Current models often do not fully integrate climate factors, population density, and hospital data to identify potential dengue hotspots. Incorporating these variables could significantly improve the accuracy of hotspot detection.
- Absence of Comprehensive Machine Learning Approaches: While machine learning models like BiLSTM, DiffSynthTab, and DiffFlow have been applied to disease diffusion in other fields, there is limited application of these models for dengue hotspot identification in Bangladesh. A more comprehensive use of these models could improve detection of high-risk areas.
- Inadequate Use of Multiple Criteria Decision-Making Techniques: Although TOPSIS and VIKOR have been employed in other fields, there is limited research on their use in dengue hotspot identification. Integrating these decision-making methods with machine learning could offer a more robust and effective approach to identifying dengue hotspots.
- Lack of Real-Time Data Integration: The majority of studies on dengue hotspots rely on historical data and static models. There is a need for models that can integrate real-time data for more effective and timely hotspot identification, which can be critical for rapid response measures.

This research aims to address these gaps by combining machine learning, geospatial analysis, and multiple criteria decision-making techniques to identify dengue hotspots more effectively.

1.3 Research Contributions

This research aims to address significant challenges in dengue hotspot identification by employing a combination of machine learning, geospatial techniques, and decision-making methods. The primary contributions of this research are as follows:

- Identification of Dengue Hotspots Using Machine Learning and Geospatial Techniques: This study introduces a novel approach for identifying dengue hotspots by integrating machine learning models such as BiLSTM, DiffSynthTab, and DiffFlow, along with traditional geographical methods such as Ordinary Least Squares (OLS) and Geographically Weighted Regression (GWR). This hybrid approach allows for more accurate identification of high-risk areas compared to conventional methods.

- Incorporation of Effective Distance for Risk Area Identification: A key contribution of this study is the introduction of an effective distance calculation method, which is used to refine the identification of dengue hotspots. By taking into account both spatial and non-spatial factors, such as proximity to certain environmental or demographic characteristics, this method helps to more accurately define the boundaries of high-risk areas for dengue transmission.

- Comprehensive Model Integration for Hotspot Detection: The research proposes a methodology that combines both classical statistical models and advanced machine learning models, utilizing climate data, population density, and hospital data to identify potential dengue hotspots. This integrated framework ensures more robust results compared to models that rely on limited datasets.

- Application of Multi-Criteria Decision-Making (MCDM) Techniques: The study employs TOPSIS and VIKOR methods from Multi-Criteria Decision Theory to classify and prioritize high-risk areas based on the results of the machine learning models. This approach enhances the decision-making process, enabling public health authorities to focus on the most vulnerable regions for preventive measures.

- Development of an Effective Hotspot Identification Tool for Public Health Authorities: By combining machine learning and geospatial analysis with MCDM techniques, this research provides a decision-support tool that enables authorities to allocate resources efficiently and take timely intervention measures to combat dengue outbreaks.

- Emphasis on Real-Time Data Integration for Improved Hotspot Detection: A key contribution of this research is the emphasis on incorporating real-time data for dengue hotspot identification. This approach is designed to provide health authorities with up-to-date information, improving the accuracy and timeliness of interventions.

- Addressing Research Gaps in Dengue Diffusion Modeling: This study fills existing gaps in dengue diffusion and hotspot identification models by offering a more comprehensive and accurate method that incorporates multiple data sources and advanced computational techniques. It extends previous research by proposing a more detailed framework for hotspot detection, with applications for epidemic management in Bangladesh

- The assessment consists of analyzing both Effective Distance (ED) model and Logistic Regression (LogReg) model against each other. Evaluation of both modeling systems occurs via performance metric analysis and involves training them on dengue risk-related data. Measurement of inter-observer reliability determines the accuracy of predictions made by both models which confirms the effectiveness of the ED model. Additional machine learning models perform verification together with model refinement to enhance the ED system's high-risk area prediction abilities. A thorough examination demonstrates that the ED model stands as a dependable and effective process for locating dangerous dengue-infected districts.

Through these contributions, this research aims to improve dengue outbreak management and provide a valuable tool for public health decision-making, ultimately contributing to better control and prevention strategies.

1.4 Organization

Chapter 2 of this thesis provides a review of current literature, examining how other researchers have employed machine learning models (MRM) and other computational approaches, along with classical geographical methods, to analyze dengue diffusion patterns. This section also covers the application of various models, including BiLSTM, Feedforward Neural Networks (FNN), DiffSynthTab, TabDDPM, and DiffFlow, in disease modeling.

Chapter 3 focuses on the data description, presenting details about the dataset used in this study, its sources, and key attributes. This chapter highlights the importance of the data collected and its relevance to the research.

Chapter 4 describes the proposed methodology, which combines AI/ML techniques, including BiLSTM, DiffSynthTab, and DiffFlow, with traditional methods like Ordinary Least Squares (OLS) and Geographically Weighted Regression (GWR). It also incorporates the concept of effective distance for risk area identification and transmission route prediction. To further enhance the accuracy, Lasso and Elastic Net regression methods are utilized within the framework.

Chapter 5 delves into the dengue hotspot identification process. This includes weight calculation, effective distance calculation, and the analysis of identified hotspots to determine the areas at the highest risk also The Effective Distance (ED) model was also validated by the Logistic Regression and Random Forest model.

Chapter 6 concludes the thesis by summarizing the key findings and contributions from Chapters 3 to 5, answering the research questions, and discussing the limitations of the study. This chapter also provides insights into the potential improvements in dengue outbreak management through more advanced computational techniques.

Chapter 2

Literature Review

In paper [18], the aim was to develop a machine learning model for regional dengue outbreak prediction. The authors of this paper tried to correlate the geographical features and population data of a region to the probability of a dengue outbreak in said region. They have developed a dataset named DengueBD which comprises data from Directorate General of Health Service of Bangladesh about daily dengue cases, weather information from Bangladesh Meteorological Department and population data from Bangladesh Bureau of Statistics. Two machine learning models, Support Vector Regression (SVR) and Multiple Linear Regression (MLR) have been used in this research. For the MLR model, the weather information was considered an independent variable and the number of dengue patients was used as a dependent variable. A cross-validation method Mean Absolute Error (MAE) was used to choose the best prediction model. 67% percent prediction accuracy has been claimed in this research using the MLR model whereas the SVR model has been 75% accurate. Drawbacks mentioned about these models are slow process for large-scale data, high memory requirement and kernel selection as only selecting the most appropriate kernel function ensures the usability of these models.

The authors of paper[27] committed to use machine learning algorithms to predict dengue cases in Bangladesh. They have used dengue related data and meteorological data from 2000 to 2013 in Bangladesh from Bangladesh Meteorological Department and Institute of Epidemiology, Disease Control and Research (IEDCR). Two features named average temperature and saturated vapor pressure have been engineered in this paper as the number of features are few. The authors tried to correlate these features to the number of dengue cases to predict dengue outbreak in the future. Machine learning algorithms used in this paper are Gradient Boosting Regression (GBR), Extreme and Light Gradient Boosting Regression (XGBR), (LGBR), Support Vector Regression (SVR) and Elastic Net Regression. And to evaluate the performance of the models used, Root-Mean-Square-Error (RMSE) has been used. The GBR models have performed better than the SVR model and Elastic Net Regression models Though, the training of the GBR models take significantly more time than SVR and Elastic Net model, GBR taking 12.23 seconds compared to SVR's 0.01 seconds. The authors have also claimed that their methodology can be used as a benchmark for future research as their prediction model has successfully predicted the peaks and downward trends of dengue cases.

Unlike the previous two papers, paper[21] tried to design a model to predict dengue from the medical information of patients alongside the meteorological data. The authors collected required data from the DengAI competition (open data of dengue illness competition: DengAI: Predicting Disease Spread (drivendata.org)), from two cities of around three to five years. For feature selection, the authors removed the humidity and precipitation as they are heavily correlated and may mess with the model accuracy. K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), Support Vector Classifier(SVC) and Gaussian Neighbor Boundaries (GNB) models are used in this research. K-fold cross validation was used to verify the performance of aforementioned models. Their performance validation shows that the RF model had an accuracy of around 87% which was the highest with DT having 72%, GNB 70%, SVC 68% and KNN 63%. Though the RF model was the most accurate, it also took significantly more time to produce results. The requirement of a significant amount of CPU power during the training phase also makes it difficult for the model to be usable on all hardware.

In paper[15], researchers build an ML model to detect whether a patient has Dengue, which gives an accuracy of nearly 100%. They have collected the needed data from the Bangladesh Public Health Association for building their machine-learning model. In total, they collected the data from 400 patients among whom 70% were dengue positive and 30% of them were dengue negative but had similar symptoms. They have chosen 12 external behaviors/ symptoms to use as attributes or features in their ML model. They have converted the symptoms into numerical or binary values to use in the process. Some of the features are High Temperature(Fever), High Plasma Rate, Headache, WBC, Haemoglobin Rate, and more. Their study employs machine learning algorithms for classification such as KNN, SVM, Naive Bayes, Logistic Regression, Random Forest, ANN, and Decision Tree. Among them, Naive Bayes and Decision Tree have the highest accuracy rate with a score of 100% followed by RF, LR, and SVM. However, ANN and KNN have the lowest accuracy rate with a score of 65% and 71% respectively. Finally, the paper doesn't talk about which constraints it has limitations or how future researchers can improve it.

Researchers are trying to predict whether a given month can be an epidemic month for dengue or not based on some given environmental data such as rainfall, humidity, and temperature in paper[20]. This research work is special in the aspect of most of the research works to detect the disease whereas it predicts a month of the year is going to be an epidemic situation for the dengue outbreak or not. So that we can be prepared beforehand. They collected the data from a research paper named "Long-term predictors of dengue outbreaks in Bangladesh: A data mining approach" from "The Infectious Disease Modelling" journal. However, they have optimized it only for Dhaka city. Their dataset has around 240 entries and each entry has 11 columns or attributes. In their data preprocessing they added a feature named "epidemic indicator" to denote the epidemic. As they are only working with the Dhaka city data they filter out the data for other cities. They have used three main machine learning approaches named SVM(Support Vector Machine), XGBOOST, and Logistic Regression(LR) to reach their goal. Among them, SVM got the height accuracy rate followed by SGBOOST and Logistic Regression respectively. They figured it out by using a confusion matrix, recall f1-score, and accuracy. As we have mentioned earlier they have worked on a pre-existing research paper, they have made it better by using SVM, which gives an accuracy rate of 97% whereas the base paper model RNN could give 94%. In future, they can work in it with a larger dataset of Bangladesh with more predictors to make it way more robust and accurate.

The authors explore various methodologies for Dengue fever prediction, moving from genetic data analysis to sensor technology and network-based predictions in paper[11]. Firstly,Caio Davi et al (2019) developed a model using support vector machines and artificial neural networks trained on immune system gene SNPs to accurately determine the severity of dengue with a high accuracy. Another approach is where Harshada Somwansh & Pramod Ganjewar (2018) proposed using Naive Bayes classification on symptoms like fever, platelets etc. for real-time dengue prediction. Sensor technology employing graphene oxide enhances detection of Dengue virus E proteins. Again, then Pengyao Ping et al (2018) developed a model using a bipartite network representation of lncRNA- disease associations that outperformed previous models in predicting associations. Ping Luo et al (2018) proposed an algorithm dgSeq that integrates protein-protein interactions, clinical RNA-seq data.These diverse approaches collectively underscore the strategies employed in the quest for accurate Dengue prediction and highlight the integrative nature of biomedical research. To Conclude,the review covers a variety of approaches leveraging machine learning, biological network analysis to tackle the prediction and detection of diseases like dengue.

In paper[25], a variety of machine learning models have been applied for dengue prediction, including support vector machines(SVM), decision trees, random forests, and neural networks(NN).This study has compared these models and found strengths and limitations of each models.In recent years, deep learning models like LSTM and CNNs have emerged as promising techniques. Here, studies highlight LSTM's ability to capture long-term dependencies in time series data, with some additional environmental factors to enhance accuracy. LSTM models have shown superior performance over other models like SVR in some studies.Additionally, adding climate and population data as inputs can improve LSTM accuracy. Attention mechanisms in LSTMs can also improve performance by identifying the most predictive input variables. One study showed LSTM with attention performed best out of CNN.Some hybrid models combining LSTM and other techniques like CNNs or neural networks have also shown good performance. Here, most studies evaluate models over short time spans of a few years.Overall, the combination of attention mechanisms with LSTM models has been specially emphasized for improved variable importance and enhanced overall predictive performance in Dengue fever forecasting.

The authors of paper[22] presents an exploration of machine learning (ML) applications in Dengue-related research and here a variety of machine learning techniques have been applied for dengue modeling and prediction, including unsupervised learning(self-organizing maps, random forests), fuzzy logic modeling, supervised learning (SVM, ANN, decision trees), and statistical modeling (regression, ARIMA).Here The studies have used different types of data features which are related to demographics, environment, climate, and dengue cases. Data availability and quality is an important consideration for this.The current focus has been on developing early warning systems using ML, with studies comparing the performance of different algorithms like SVM, ANN and random forests.Additionally, the review emphasizes the potential of ML and artificial intelligence advancements in revolutionizing healthcare particularly in predicting diseases like Dengue.In summary, the review provides a comprehensive overview of the current state of ML applications in Dengue research,giving importance to the key themes and gaps, and offers insights into the potential future directions of research in this domain.

In paper[10], the WEKA device is used on this paper to are expecting dengue outbreak. The authors of the paper evolved and skilled a version over a dataset pattern. Additionally, seven well-known machine gaining knowledge of techniques had been implemented and eight parameters were used for performance assessment. They have used Receiver Operating Characteristic (ROC) to estimate the performance of the class models over all possible thresholds. After comparing the results,It has been concluded that the LogitBoost ensemble model is the topmost overall performance classifier approach that has reached a type accuracy of ninety two% with sensitivity and specificity of 90 and 94% respectively and ROC location=zero.967, and had the lowest errors rate. In addition to that, they have got compared the accuracy price of our analysis with other posted results. Based at the comparative evaluation end result using the LogitBoost ensemble version in addition to the Random forest classifier used by Fathima et al, (2015) result concluded that the ensemble version plays higher than the person classifier. Limitations of this studies referred to that a unmarried getting to know set of rules cannot be the great and at most appropriate with the complete area of utility. It may be that the computing cost and processing time can growth due to the ensemble model.

The paper[14] reports on a study that uses statistical machine learning algorithms to automate the system of predicting the Case Fatality Grade (CFG) of Dengue fever. As an extension of telemedicine, the work pursuits to create fundamental statistical classifiers that may expect Dengue CFG at a level of accuracy that is commensurate with that of specialized seed health practitioner. The authors contributed with statistical modeling and paper writing similarly to collecting and device-mastering scientific information. The study uses information that turned into collected for a hundred patients in diverse Indian hospitals and clinics among 2016 and 2019. In order to automate the human gaining knowledge of protocol by means of imitating functional outliers, the look at makes use of a combination of statistical processors from information studying to hazard control. The authors also communicate about how scientific judgment and patient effects may be tormented by automating the Dengue CFG prediction technique thru statistical gadget learning. The paper additionally outlines future work, recognising the observe's boundaries and suggesting that the SML classifier-based prediction set of rules be extended to account for stochastic fluctuations in a clinical choice-making system via fusing scientific 2d-first-rate diagnostics with Bayesian inferencing. The facts profile's primary innovation is its distinctiveness and longitudinal scope. The have a look at makes use of a variety of statistical processors, from hazard evaluation to data analyzing.

In paper[13], researchers build an ML model to predict dengue fever type at an earlier stage. The writers initially gathered information on 209 individuals with dengue from two main medical institutions in Bangladesh - Dhaka and Chittagong Medical Colleges. The information consisted of an equal number of regular fever (75), regular dengue fever (74), and serious dengue fever (60) cases. The information consisted of details about the patients like their age, gender, and background, their past health records, tests done in the laboratory, and their symptoms in 23 different aspects. This unprocessed data was prepared by dealing with absent values, standardization, and characteristic selection. Next, the dataset was divided into two parts with a ratio of 70:30(70 training ,30 test). The training set consisted of 146 cases, while the test set contained 63 cases. Two supervised machine learning methods were examined -decision tree (DT) employing CART and Gini impurity, and random forest (RF) utilizing bootstrap aggregation. During the process of 10-fold cross validation, the decision tree (DT) model demonstrated superior performance compared to the random forest (RF) model. The DT model achieved an accuracy of 79% on the test set, while the RF model only achieved 74%. The decision tree method successfully forecasted the type of dengue fever based on the patient dataset containing 23 characteristics.

In the paper[17], researchers build an ML method known as the PFDM to promptly and precisely detect dengue illness. The writers gathered an actual dataset consisting of 347 records of people with dengue from a state-run hospital in Lahore, Pakistan. The set of data had 7 main characteristics/factors capturing essential clinical and blood information needed for diagnosing dengue, including sex, years, level of hemoglobin, count of white blood cells, level of hematocrit, count of platelets, and outcomes of IgG/IgM/NS1 antigen tests. After collecting data, the information was processed by gathering data, increasing the number of samples, and dividing it into training (242 records) and testing (105 records) sets with a proportion of 70:30. Two supervised machine learning models, namely Artificial Neural Network (ANN) and Support Vector Machine (SVM), were provided with training data to undergo training. The artificial neural network (ANN) model used a 6-level structure with Bayesian regularization, while the support vector machine (SVM) created the best hyperplanes to distinguish between the two groups of dengue cases: positive and negative. After the training process, the forecasts from the artificial neural network (ANN) and support vector machine (SVM) were combined using a decision system based on fuzzy logic. This combination resulted in the final outcomes of the PFDM model. Thorough examination of the 105 samples showed that the suggested PFDM model obtained a great precision of 96.19% in categorizing dengue cases, surpassing the individual ANN (95.24%) and SVM (93.33%) models. The notable enhancement in performance shows the advantages of combining artificial neural network and support vector machine predictions using fuzzy logic, to create a precise machine learning-based system for detecting dengue early on with the help of patient's clinical and lab information.

Lastly, on paper[19], a study conducted by Khan and colleagues (2022) concentrates on creating computerized learning models to anticipate dengue outbreaks in Bangladesh at an early stage. The writers gathered an actual set of information from 206 individuals in various medical facilities in Noakhali and Dhaka, Bangladesh. The well-rounded dataset collected 25 characteristics such as medical symptoms, basic population information, and lab examinations that are important for determining if someone has dengue fever. The information was split into three parts: a training set (144), a validation set (31), and a testing set (31), following a ratio of 70:15:15. Three supervised artificial neural network methods - Scaled Conjugate Gradient (SCG), Levenberg-Marquardt (LM), and Learning Vector Quantization (LVQ) - were utilized and contrasted. The language model using a hidden layer architecture obtained the best accuracy rate of 97.3% on the test dataset, surpassing SCG (87.1%) and LVQ (90.3%) by a large margin. The writers also created a Decision Tree (DT) classifier, which had a training accuracy of 98.92% but only 90% accuracy when tested. In general, the Levenberg-Marquardt neural network was discovered to be the most efficient machine learning method for early and precise estimation of dengue epidemics in Bangladesh using the 25-characteristic patient dataset. The study shows the effectiveness of guided profound learning techniques such as language model for foreseeing the occurrence of disease outbreaks using clinical and demographic information.

As for diffusion pattern models, in paper [26], the authors have proposed a model that would use machine learning to detect covid-19 hotspots. The diffusion model they proposed in the paper works in two modes, one where lag is considered and another where there is considered to be no lag. Here, lag refers to the time it takes to attain maximum value and no lag means instantaneous spread of the virus.
Here, n refers to the length of the dataset, e is Euler's number, β is the number of cases changing from day to day when lag is considered and θ is the same value but when lag is not involved. As for γ, it is calculated using the following formula, where α is the learning rate of the model. The performance comparison between the proposed model and other existing machine learning models is represented using various graphs in the paper, and oftentimes the proposed model has shown error rate 10% less than models like SVM, LR, CNN.

The authors of paper[7] used climatic and vector data from year 1996 to 2011 to correlated spatial and temporal diffusion patterns of dengue outbreak to these data in Belo Horizonte, Brazil which is very close to the amazon rainforest and a ideal breeding ground for aedes mosquitoes. They have used data from various primary sources including municipal surveillance data, home addresses of patients and even aedes larvae. For meteorological data, they collected necessary information from the Brazilian Meteorological Institute (INMET). The authors have used geographical information systems (GIS) and Kernel density estimation to find and predict the diffusion of dengue across the area. They have also found that humid and rainy weather correlated to higher numbers of dengue cases and also these epidemics arise usually in five distinct periods of time over the years. Another finding from this paper is that dengue hotspots often occur in areas with lower elevation as this can lead to higher amounts of stagnant water. Using the data on aedes larvae, they found that the amount of larvae directly correlated to the number of patients next month. The authors finally suggested employing the dengue spread countermeasures on the predicted and identified hotspots rather than brute force approach in dengue contamination across the whole city.

In paper[8], the authors aimed to analyze the spatial and temporal distribution of dengue, basically the dengue spread across Swat, Pakistan. They collected patient data from various hospitals across the area and used space-time scans and ring maps as spatio-temporal data for the research. The authors have implemented two methods over the data they have collected to do the analysis, Ordinary least squares (OLS) and Geographically weighted regression (GWR). Their finding was that the densely populated urban areas near southern parts of Swat were highly vulnerable to dengue. As they have used a huge amount of historical data, they could also identify historical dengue clusters, among which the cluster from August to October in 2023 was the largest. They also found that the younger demographic is more likely to become infected as over 55% of the patients historically have been under the age of 40. Although, unlike the paper before, this paper found a negative correlation between elevation and dengue cases. In their research, the GWR model performed better in all test scenarios and had higher R-squared values.

Lastly in paper[5], the researchers tried to find the spatio-temporal dengue diffusion pattern in Chachoengsao Province, Thailand, from 1999 to 2007. They had collected data from various villages in the area for patient data and used meteorological data from the Thai Meteorological Department. From their analysis, they found that from temporal analysis, the months May to September show high correlation between the number of dengue patients and weather factors such as rainfall and humidity. The authors also applied Empirical Bayes Smoothing (EBS) on the data along with Local Indicator of Spatial Association (LISA) and Getis-Ord Gi* statistics. The new finding from this paper is that the correlation between weather factors and the dengue cases have a month of lag between them, meaning higher rainfall or humidity correlated to higher dengue cases the next month. The authors also used ArcGIS, GeoDa, and SatScan and found similar results to other papers that the dengue hotspots normally occur in densely populated regions and urban areas.

Chapter 3

Dataset

3.1 Dataset Creation

For this research, we had to collect patient data for all 64 districts of Bangladesh which include the number of total admitted patients, released patients, and the number of confirmed deaths in each district for each month from September 2019 to October 2024.

However, due to the COVID-19 pandemic, the patient data from February 2020 to December 2021 could not be collected from any sources as the medical sector as a whole was more concentrated on the coronavirus pandemic. The patient data has been collected from the Directorate General of Health Services of Ministry of Health and Family Welfare, Bangladesh. Daily press reports containing the number of patients in each district are uploaded to their official webSite.

For meteorological data, we used the Visual Crossing Weather API for weather data (rainfall, humidity, maximum and minimum temperature) for all 64 districts in each month of the dataset. The data could not be collected from the Bangladesh Meteorological Department due to it not having data from all 64 districts for our required time frame.

Lastly, we collected the population data from the Bangladesh Bureau of Statistics. Using the data from the “Population and Housing Census 2022” and “Population and Housing Census 2011”, including the total population, area, and population growth rate from 2011 to 2022, we calculated the population density of each of the 64 districts for all months in the dataset.

3.2 Data Description

The dataset contains comprehensive information on dengue cases, climatic conditions, and geographical data for 64 districts in Bangladesh between 2019(September) and 2024(October). With 2,496 entries and 14 columns, it contains important health indicators like the total number of patients admitted, released, deaths, and current patients under care; climate data including maximum and minimum temperatures, rainfall, and humidity levels. Each district also includes geographic identifiers like latitude, longitude, and population density. The dataset provides insightful information about the connection between environmental factors and dengue's spread.

Year	Month	District	Total Admitted	Released	Death	Current Patient	Max Temp	Min Temp	Rainfall	Humidity	Latitude	Longitude	Population Density
2019	September	Dhaka	6545	8670	23	563	36	24.5	192.6	79.06	23.8103	90.4125	9838.907981
2019	September	Faridpur	357	374	0	17	36.1	24.6	171.6	79.65	23.6088	89.8172	1055.263731
2019	September	Gazipur	131	158	0	8	36	24.6	215.4	78.7	24.0034	90.4118	2670.14672
2019	September	Gopalganj	141	150	0	18	35.3	24.1	181.5	86.12	23.0163	89.8268	887.6062286
2019	September	Jamalpur	69	79	0	6	36.2	23.6	216.7	85.99	24.9226	89.9456	1195.541769
2019	September	Kishoreganj	164	191	0	7	35.5	24.4	253.5	80.48	24.4428	90.745	1218.668626
2019	September	Madaripur	138	156	0	8	36	24.1	192.6	80.45	23.1663	90.2076	1155.122731
2019	September	Manikganj	610	649	0	57	36	24.7	171.8	79.13	23.857	90.009	1129.653845
2019	September	Munshiganj	100	121	0	4	36	24.7	187.5	79.15	23.5476	90.5281	1622.109451
2019	September	Mymensingh	224	292	0	23	37.4	24.1	251.7	86.27	24.747	90.4152	1338.092472
2019	September	Narayanganj	124	144	0	12	36	24.7	192.6	78.97	23.6112	90.497	5468.281959
2019	September	Narsingdi	123	133	0	16	35.7	24.6	246.2	79.52	23.9107	90.7111	2230.664388
2019	September	Netrokona	14	14	0	1	37.5	24.1	257.4	85.73	24.8833	90.734	851.899778
2019	September	Rajbari	79	88	0	4	36	24.7	177.8	79.73	23.7081	89.635	1088.075472
2019	September	Shariatpur	144	156	0	9	36	24.3	192.6	80.23	23.2381	90.3583	1106.191991
2019	September	Sherpur	25	35	0	2	36.2	23.6	216.7	86.08	25.0216	90.0191	1108.919039
2019	September	Tangail	204	220	0	22	36	23.7	180.8	79.47	24.25	89.95	1186.414051
2019	September	Bogra	178	225	0	16	36.1	23.6	290.4	86.19	24.857	89.373	1289.76174
2019	September	Joypurhat	14	16	0	1	36	23.4	312.7	86.54	25.0946	89.0247	958.339606
2019	September	Naogaon	30	31	0	1	36.1	23.6	307.7	86.36	24.8043	88.939	1156.571067
2019	September	Natore	29	36	0	2	36	23.9	294.4	86.02	24.4061	88.9836	981.9214692
2019	September	Nawabganj	69	90	0	5	36	24.7	169.5	79.09	24.5865	88.2647	1074.951342
2019	September	Pabna	197	216	0	10	36	24.3	199.9	85.51	24.008	89.24	1210.282632
2019	September	Rajshahi	198	207	0	13	36.1	23.7	288.9	86.09	24.3738	88.6014	1195.968057
2019	September	Sirajgonj	307	337	0	18	36.1	24.1	220.6	85.64	24.4538	89.702	1404.30212

Table 3.1: Sample Dataset

Column Name	**Data Type**	**Description**
Year	`int64`	Year of the recorded data.
Month	`object`	Month of the recorded data.
District	`object`	Name of the district where data was collected.
Total_admitted	`int64`	Total number of patients admitted.
Released	`int64`	Total number of patients released.
Death	`int64`	Total number of deaths recorded.
Current_patient	`int64`	Number of patients currently under care.
Max_temp	`float64`	Maximum temperature recorded (in Celsius).
Min_temp	`float64`	Minimum temperature recorded (in Celsius).
Rainfall	`float64`	Amount of rainfall recorded (in millimeters).
Humidity	`float64`	Average humidity recorded (in percentage).
Latitude	`float64`	Latitude of the district.
Longitude	`float64`	Longitude of the district.
Population_density	`float64`	Population density of the district (people per square kilometer).

Table 3.2: Description of the dataset columns

The dataset contains 2,496 rows and 14 columns in total of 34,944 data points. Key columns are: Year, Month, District, Total Admitted, Released, Death, Current Patient, Max temp (C), Min temp (C), Rainfall (mm), Humidity (%), Latitude, Longitude, and Population Density .This data enables the identification of climatic patterns and their impact on dengue's spread across the district.

3.3 Exploratory Data Analysis (EDA)

3.3.1 Skewness:

Skewness measures the asymmetry of data, where a positive value denotes a right-tailed distribution and a negative value denotes a left-tailed one. Here, High positive skewness is seen in columns such as Total_admitted, Released, Death, Current_patient, TOPSIS_Closeness and VIKOR_Value, which show that the majority of values are grouped at the lower end with a few extremely high values.Conversely, VIKOR_Value has a notable negative skewness, indicating a longer left tail and a concentration of higher values. With skewness near 0, columns like Max_temp, Min_temp, Latitude, and Longitude are almost symmetrical.

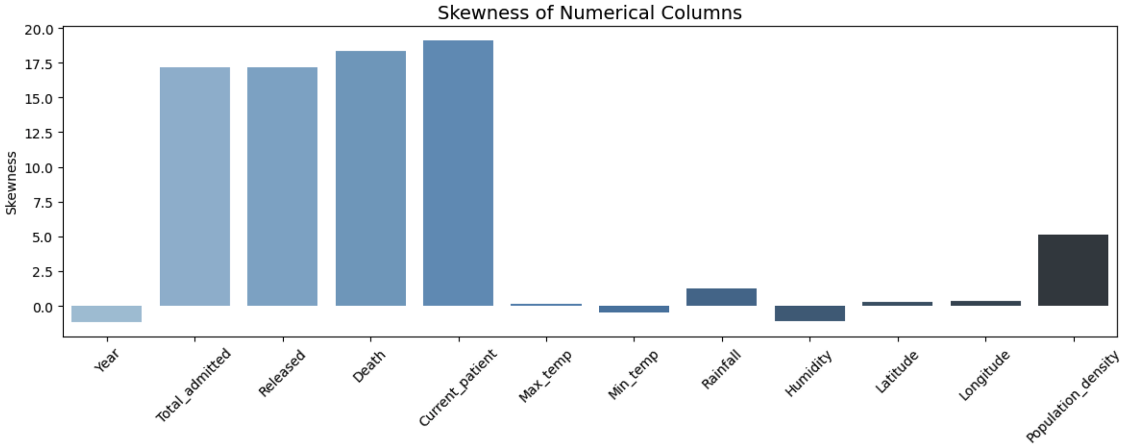

Figure 3.1: Skewness

3.3.2 Kurtosis:

Kurtosis measures the ”tailedness” of a distribution, revealing that health-related columns such as Death, Total_admitted, and Current_patient have extremely high kurtosis, indicating the existence of heavy tails and significant outliers. On the other hand, the low kurtosis of Max_temp, Min_temp, and geographic data such as Latitude and Longitude suggests lighter tails and more consistent distributions.Remarkably, kurtosis values of -1.2 are displayed by TOPSIS_Rank and VIKOR_Rank, suggesting a uniform or flat distribution.Overall, the data shows that patient-related data is very varied, with some districts or months witnessing significant dengue incidence spikes, whereas the climatic factors are relatively consistent across districts.

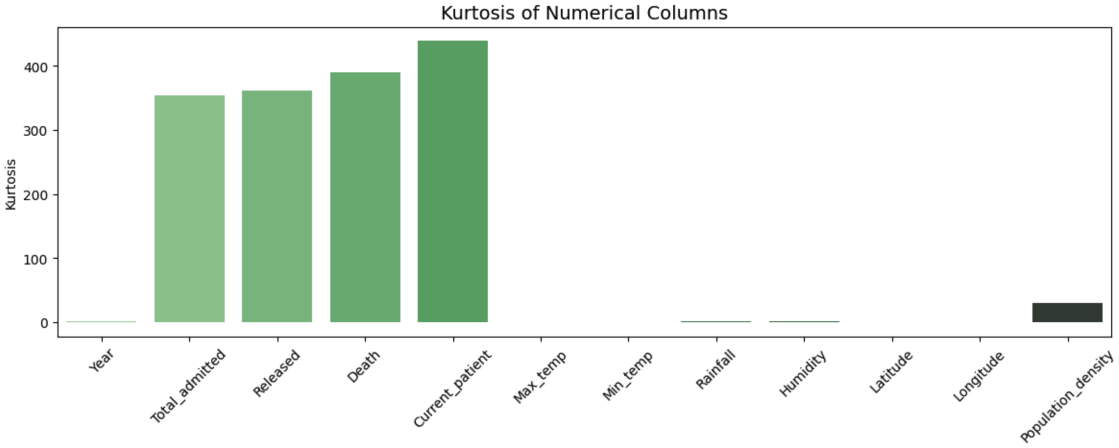

Figure 3.2: Kurtosis

Variable	**Skewness**	**Kurtosis**
Year	-1.172925	0.726487
Total_admitted	17.139464	353.582445
Released	17.201574	361.608935
Death	18.349915	390.576276
Current_patient	19.107466	438.858004
Max_temp	0.157067	-0.127016
Min_temp	-0.488462	-1.003144
Rainfall	1.244030	1.598366
Humidity	-1.119816	1.253682
Latitude	0.257889	-0.645665
Longitude	0.331975	-0.716326
Population_density	5.102688	29.135579

Table 3.3: Skewness and Kurtosis.

The analysis reveals a small number of exceptional cases are responsible for the considerable variation in dengue-related data between districts and time periods. Patient-related columns, including Total_admitted, Released, Death, and Current_patient, show high positive skewness and kurtosis, suggesting that outbreaks are concentrated in particular areas or times.On the other hand, low skewness and kurtosis are displayed by climatic variables such as Max_temp, Min_temp, Latitude, and Longitude, suggesting uniform and regularly distributed conditions throughout districts. Extreme kurtosis in patient data draws attention to notable outliers, which most likely indicate serious outbreaks or abnormalities in dengue incidence. This underscores the necessity of focused intervention in areas that are impacted.

3.3.3 Rolling mean and Residuals:

Using a window size of three, this analysis computes the rolling mean and residuals for important numerical columns such as Total_admitted, Released, Death, Current_patient, Max_temp, Min_temp, Rainfall, and Humidity. While the residuals draw attention to short-term variances, the rolling mean smoothes the data and reveals long-term patterns.The rolling mean indicates steady seasonal trends in environmental variables and stabilization in patient-related data. While smaller residuals in temperature data indicate consistent changes, larger residuals in patient data represent abrupt outbreaks.

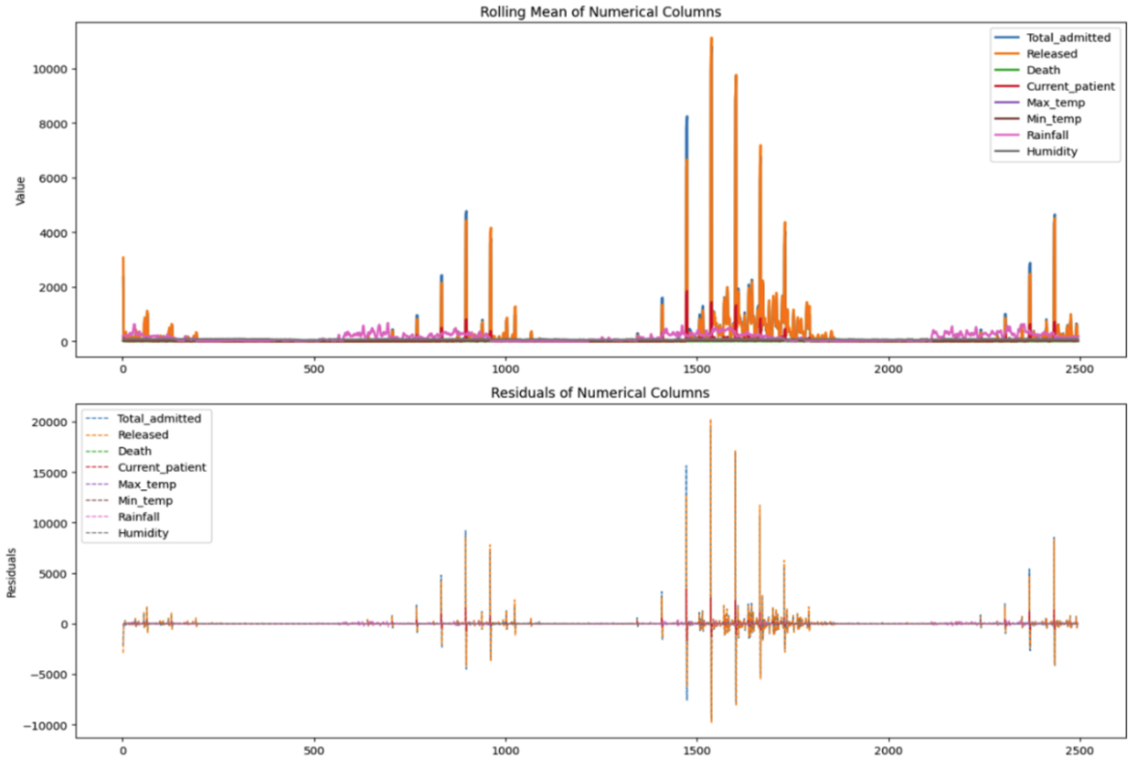

Figure 3.3: Rolling Mean and Residuals

3.4 Data Visualization

Correlation Matrix:

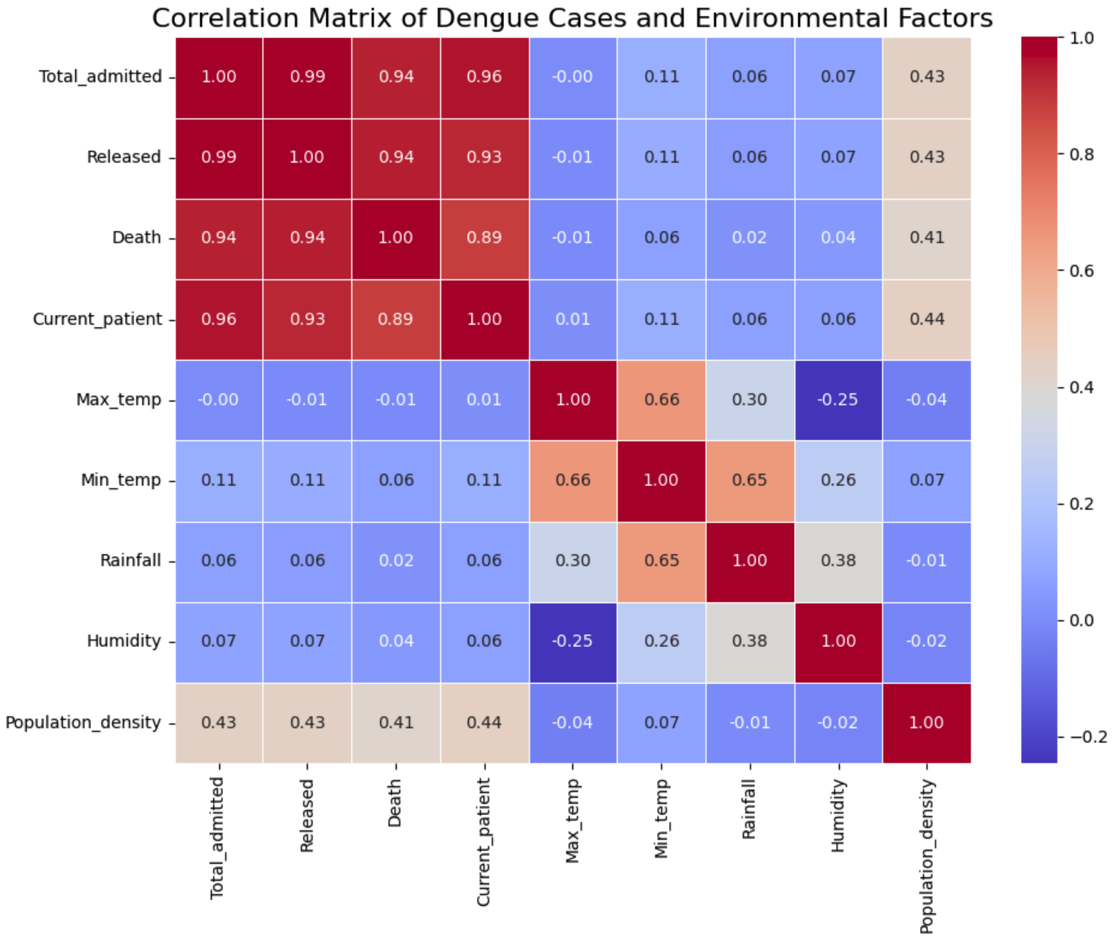

Figure 3.4: Correlation Matrix

The correlation matrix shows strong positive relationships among patient-related variables such as Total_admitted, Released, Death, and Current_patient, suggesting that as one of these increases, the others tend to increase as well.While Min_temp and Rainfall have moderately strong correlations with Current_patient and Released, showing that temperature and rainfall may affect dengue case trends, Max_temp has weak relationships with patient data, suggesting negligible impact on patient counts.Additionally, there are moderate connections between population density and patient characteristics, suggesting that dengue cases may be more common in places with larger population densities. Overall, the data show that patient counts are strongly correlated, although the effects of population density and climate are less pronounced.

Total Admitted Patients by District:

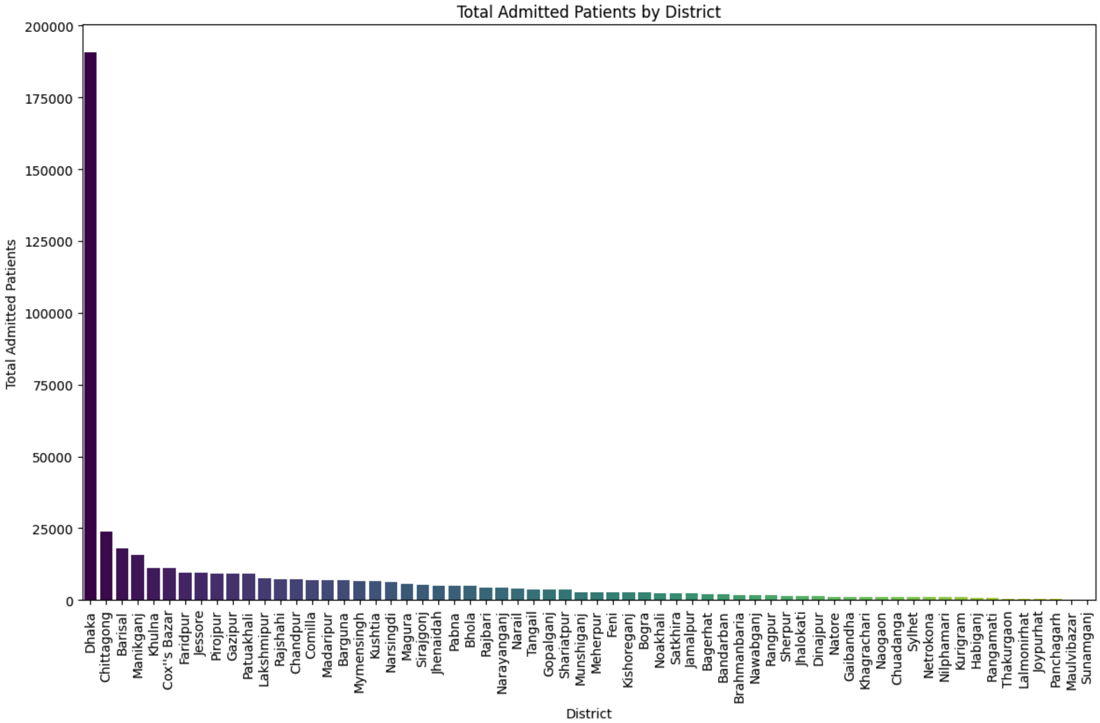

Figure 3.5: Total Admitted Patients by District

The bar graph showing the total number of admitted patients by district. The y-axis displays the total number of patients admitted, while the x-axis lists the districts. The number of patients admitted to various districts is shown visually in the graph, with the Dhaka district having the most admitted patients in comparison to the other districts.

Total Deaths by District:

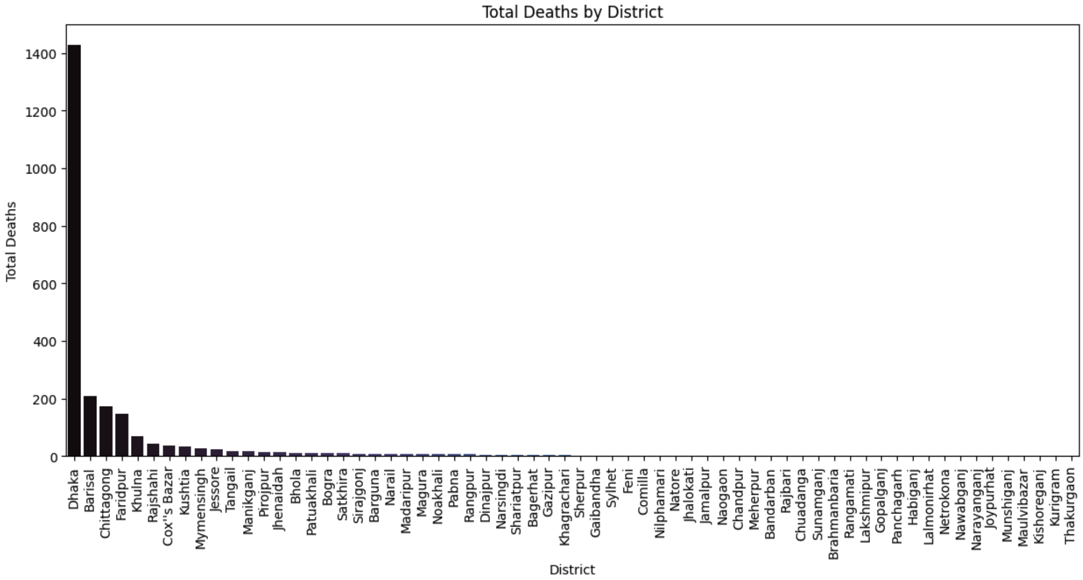

Figure 3.6: Total Deaths by District

The bar graph shows the total deaths by district. The y-axis displays the total number of deaths, while the x-axis lists the districts. The number of deaths due to dengue disease in various districts is shown visually in the graph, with the Dhaka district having the most deaths in comparison to the other districts.

Current Admitted Patients by District:

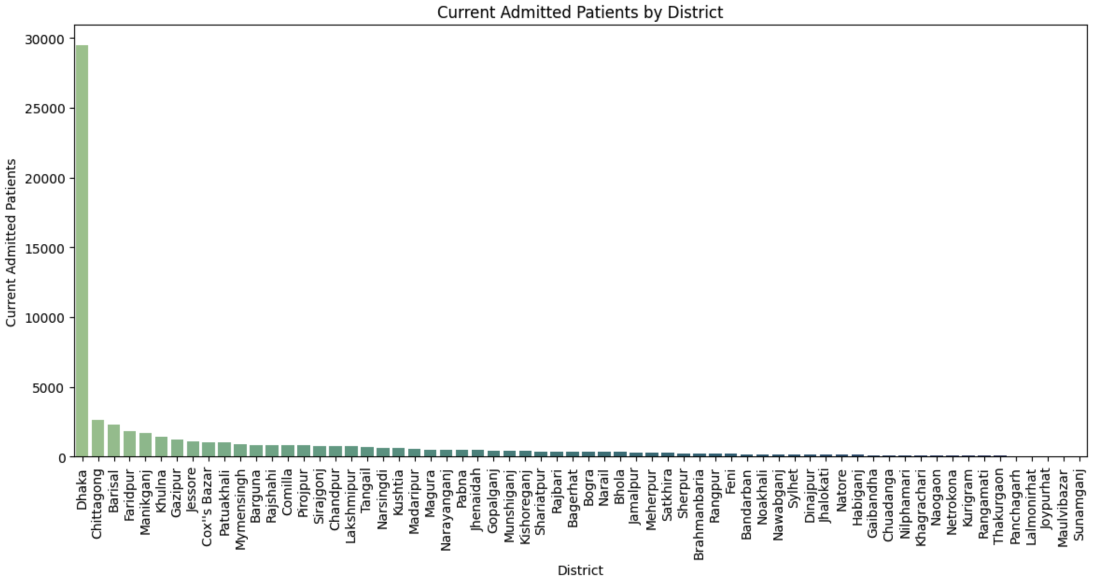

Figure 3.7: Current Admitted Patients by District

The number of Current Admitted Patients in various districts is shown visually in the graph, with the Dhaka district having the most current Admitted Patient in comparison to the other districts.

Total Released Patients by District:

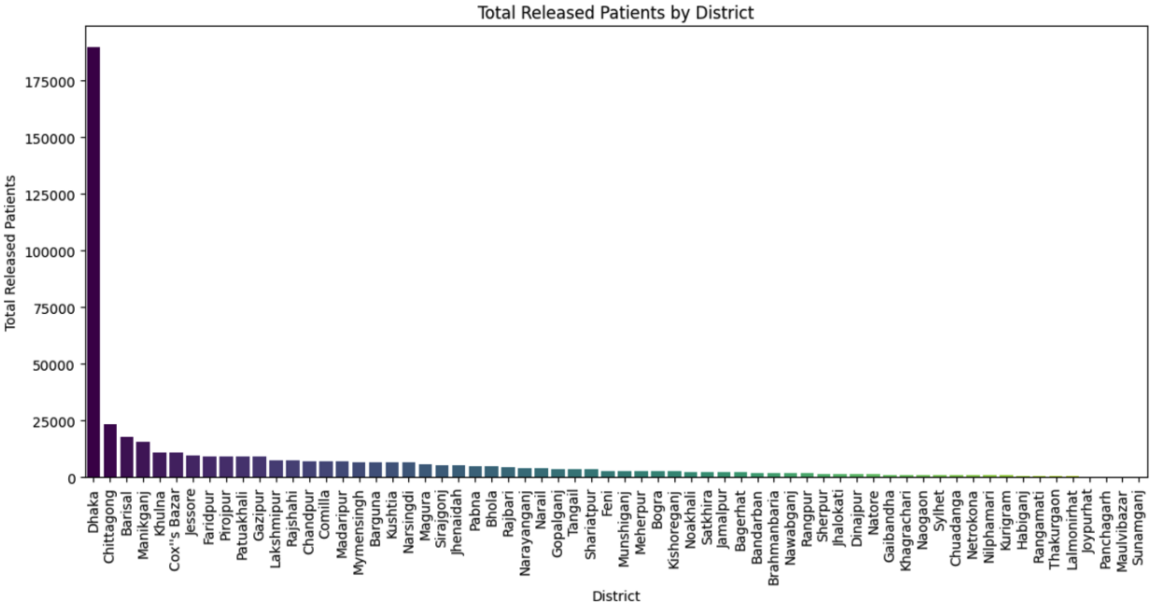

Figure 3.8: Total Released Patients by District

The number of released patients in various districts is shown visually in the graph, with the Dhaka district having the most released patients in comparison to the other districts.

Patient Categories in top 5 District:

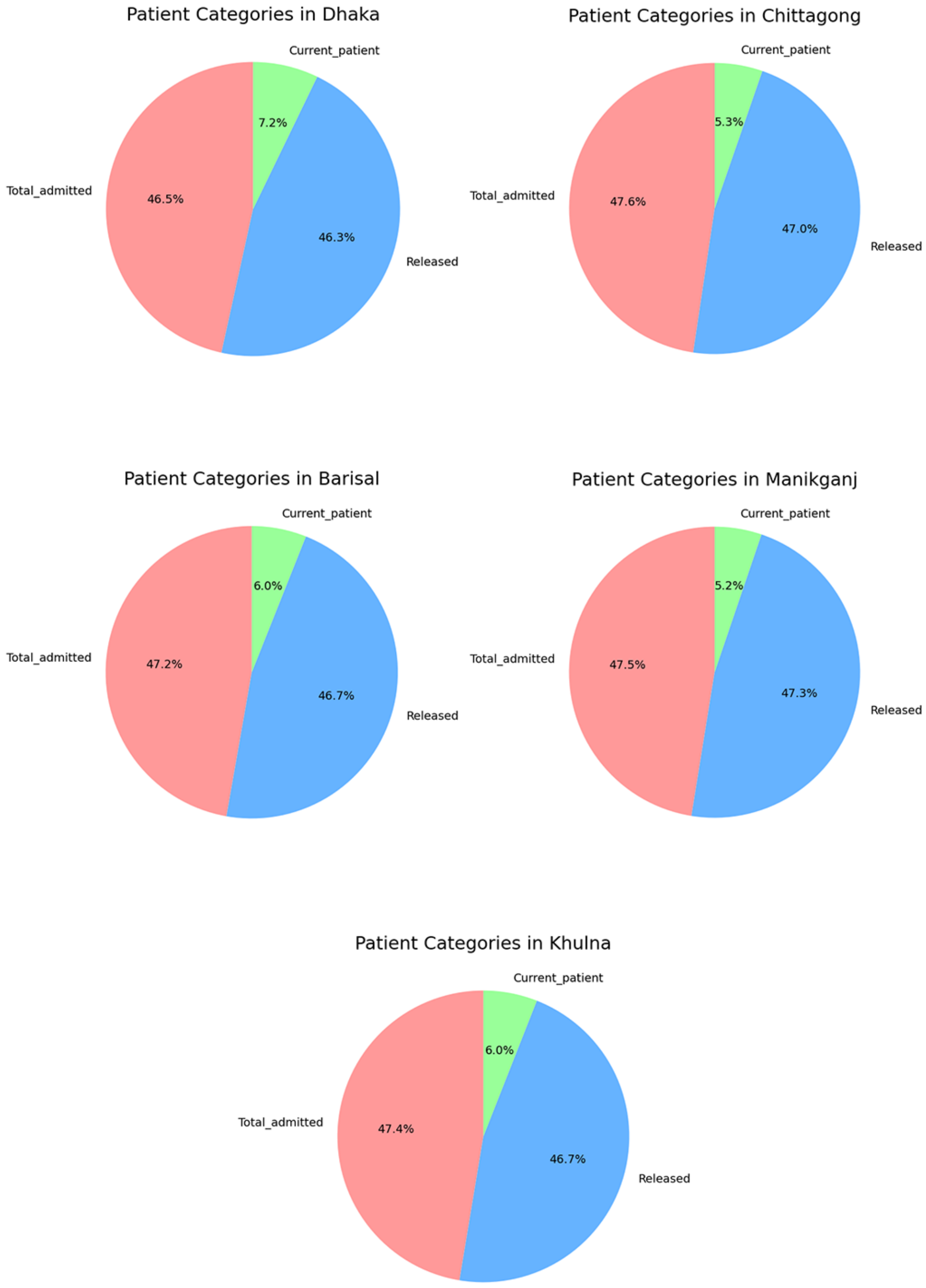

Figure 3.9: Patient Categories in top 5 District

Top 5 Districts by Deaths in 2024:

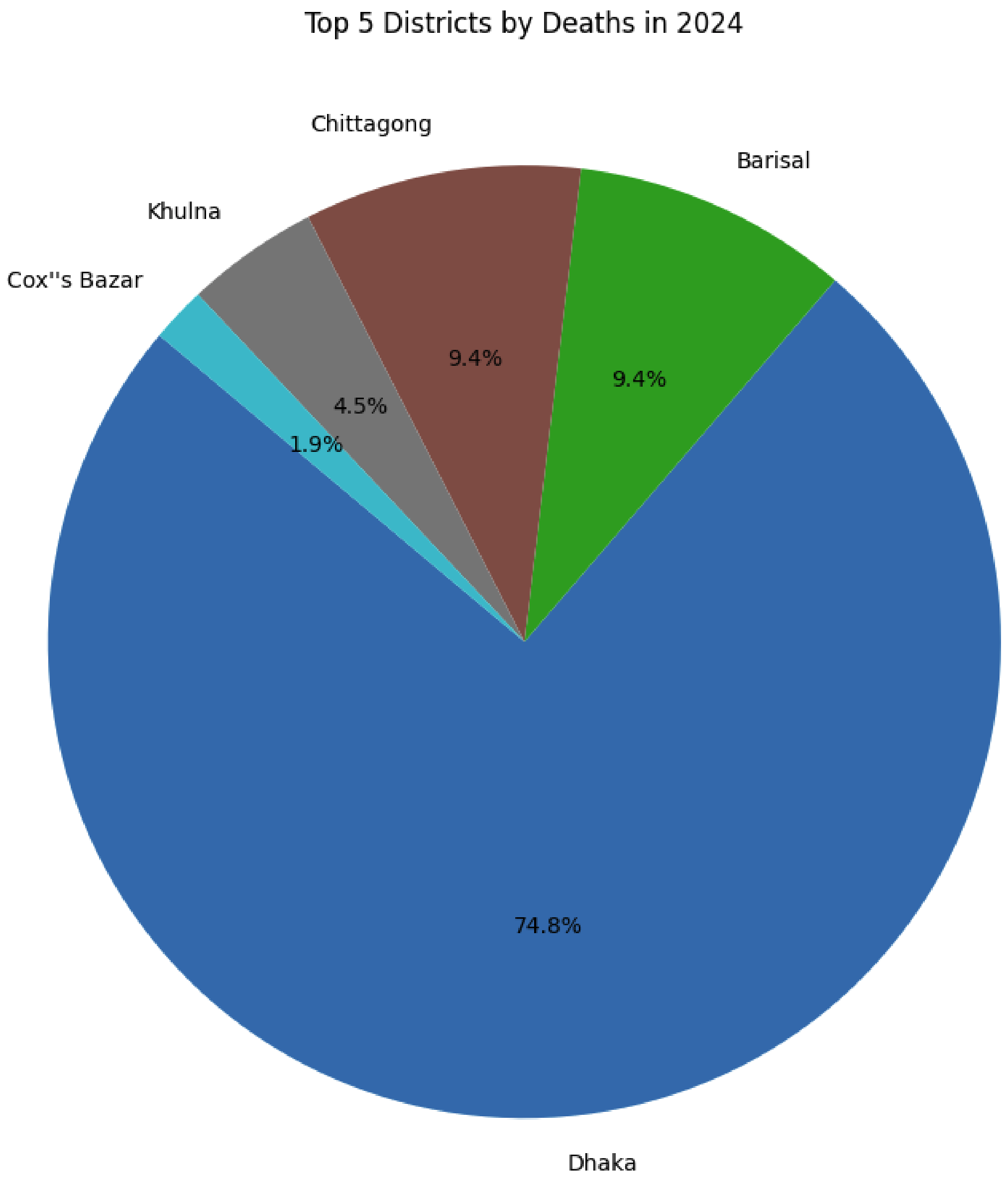

Figure 3.10: Top 5 Districts by Deaths in 2024

The top 5 districts by number of deaths in 2024 are displayed in the pie graphic. With 74.8% of all deaths, Dhaka comprises the greatest portion of the pie. Chittagong (9.5%), Barisal (9.5%), Khulna (4.5%), and Cox's Bazar (1.9%) are the next largest portions.

Max and Min Temperature Trends Over the Years:

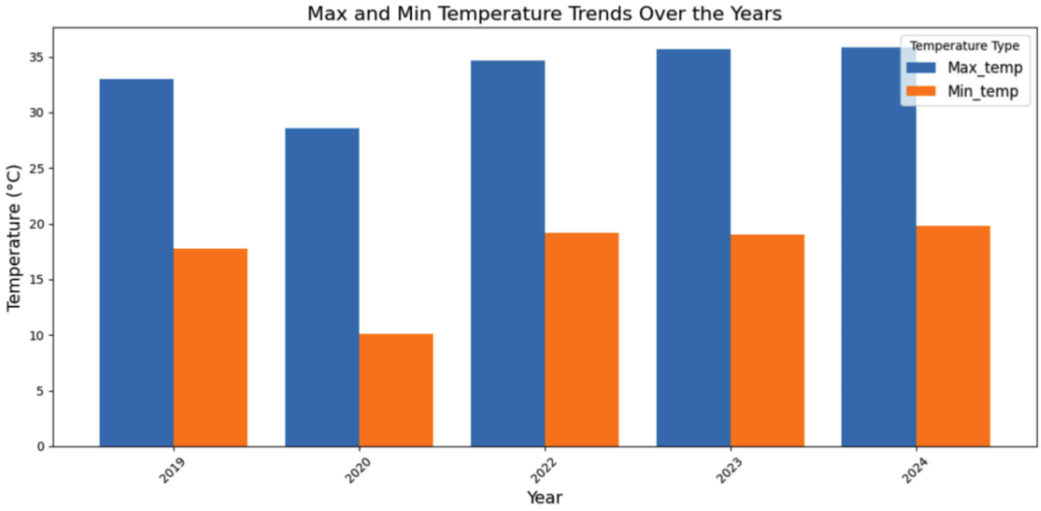

Figure 3.11: Max and Min Temperature Trends Over the Years

The graph displays the trends in the highest and lowest temperatures over time. With the maximum temperature rising from about 32°C in 2019 to 34°C in 2024 and the minimum temperature rising from around 17°C to 20°C in the same time frame, it shows that both the maximum and minimum temperatures have been rising steadily.

Rainfall and Humidity Trends Over the Years:

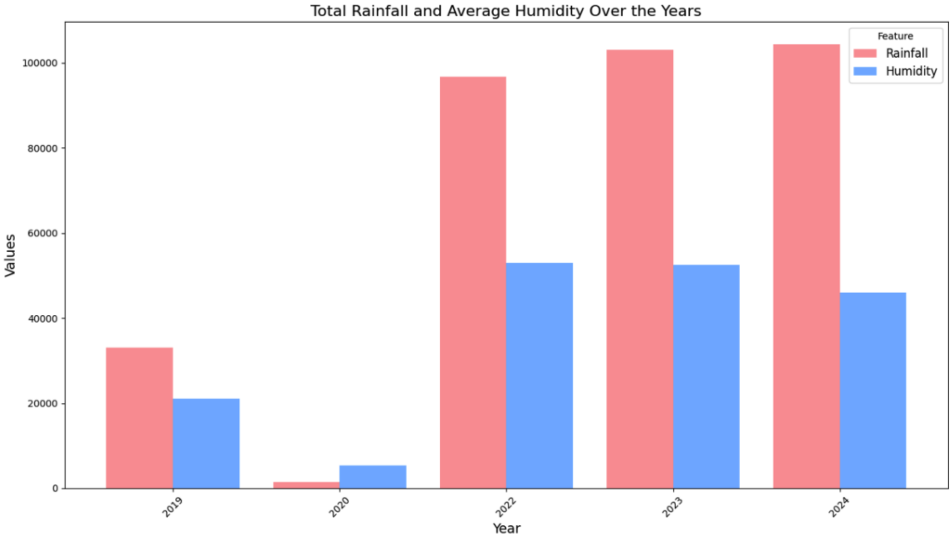

Figure 3.12: Rainfall and Humidity Trends Over the Years

The graph shows the trends in total rainfall and average humidity over the years. It indicates that both rainfall and humidity have fluctuated, with significant increases in 2022 and 2023 followed by a decrease in 2024.

Chapter 4

Methodology

Our proposal of detecting dengue diffusion patterns & identifying hotspots completely depends on the dataset based on different hospital records from all 64 districts of Bangladesh. There are multiple machine learning algorithms, but in our proposed model, We will use a combination of machine learning and geospatial models to compare their performance on such datasets. Afterwards, we will use an effective distance algorithm combined with TOPSIS and VIKOR for hotspot detection and prediction. As a result, implementing this model in real-world scenarios will help authorities control the dengue virus and raise awareness.

For this research, we first collected the necessary data from various sources. After compiling the data into a new and unique dataset, we proceeded with preprocessing to align it with our use case. In the preprocessing stage, coordinate values for each district were added to the dataset to facilitate models such as the Effective Distance Algorithm. Following data preprocessing and analysis, the dataset was visualized using multiple graphs and charts. Subsequently, TOPSIS and VIKOR were applied to the dataset, focusing on key target variables: Total Admitted, Released, and Death. Using entropy-based calculations derived from the ranked columns, we determined the required weights for these variables. Concurrently, multiple machine learning and diffusion models were implemented on the dataset. Based on the best-performing model, the weights for these variables were recalculated. Subsequently, a weighted average was computed using the values from TOPSIS and VIKOR along with the weights derived from the best-performing model. These final weights were then utilized in the Effective Distance Model to identify dengue hotspots.

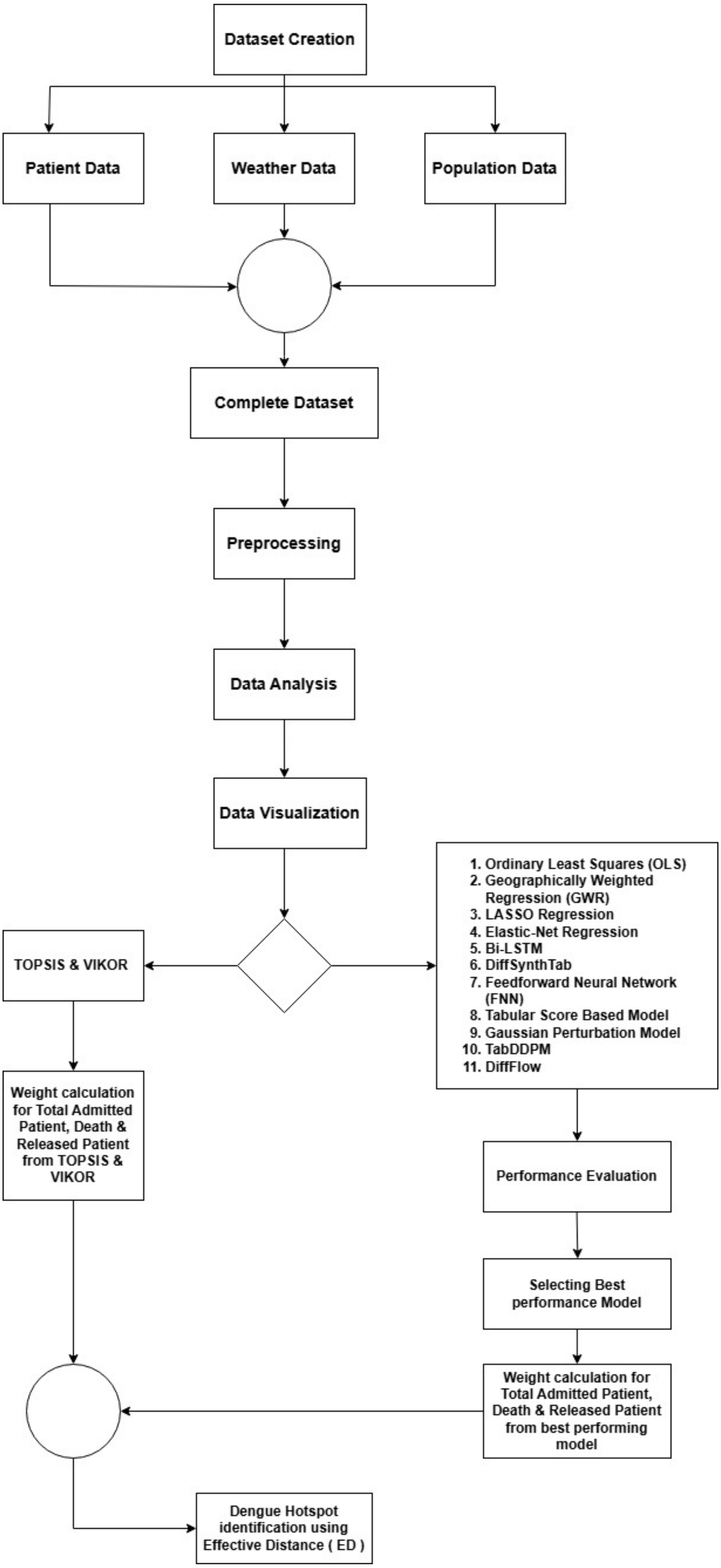
Dataset Creation
Patient Data
Weather Data
Population Data
Complete Dataset
Preprocessing
Data Analysis
Data Visualization
TOPSIS & VIKOR
1. Ordinary Least Squares (OLS)
2. Geographically Weighted Regression (GWR)
3. LASSO Regression
4. Elastic-Net Regression
5. Bi-LSTM
6. DiffSynthTab
7. Feedforward Neural Network (FNN)
8. Tabular Score Based Model
9. Gaussian Perturbation Model
10. TabDDPM
11. DiffFlow
Weight calculation for Total Admitted Patient, Death & Released Patient from TOPSIS & VIKOR
Performance Evaluation
Selecting Best performance Model
Weight calculation for Total Admitted Patient, Death & Released Patient from best performing model
Dengue Hotspot identification using Effective Distance (ED)

Figure 4.1: Top Level Overview

4.1 Model Specification

As dengue is not contagious, we cannot analyze and understand dengue diffusion just using the patient data, as the disease spreads through infected aedes mosquitoes rather than from human to human. Thus, we need to understand the relation between the surge of patients and the change of natural indicators to understand how and when the disease spreads. For example, rainfall, temperature and various other measurements can determine the number of aedes mosquitoes directly or indirectly. Therefore, change of these variables in an area can also cause the change in numbers of patients in that area.
So, we can use diffusion model algorithms and machine learning algorithms to determine the approximate number of patients and understand the diffusion and spread of dengue fever. As we will not be getting a predetermined output from our model, our goal cannot be achieved by using classification analysis, rather we have to use regression analysis in our case. For this research, various algorithms have been used to simulate the dengue diffusion and some machine learning algorithms have also been used to compare our findings. Details about these algorithms have been discussed below.

4.1.1 Ordinary Least Square Algorithm

The formula of the ordinary least square algorithm is as follows [3]

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon \tag{4.1}$$

Here, β_0 corresponds to the intercept of the model. If the model is running on a dataset where there are j number of features that the model trains on, X_j then refers to the j th variable where j can range from 1 to p and ϵ denotes the random error of the model.

This model works by trying to minimize the sum of square difference between the prediction and observation. The prediction vector is derived using the following formula:

$$y^* = X\beta = X(X^{'}DX) - 1X^{'}Dy \tag{4.2}$$

Here, y is the vector of observed value for the dependent variable. X and D are both matrices where X denotes a matrix with a vector of 1s followed by a matrix of explanatory variables. And D is a matrix with the weight values on its diagonal. The error value has a variance of which is calculated using the following formula:

$$\sigma^2 = 1/(W - P^*) \sum_{i=1}^{n} wi(yi - y^*i) \tag{4.3}$$

4.1.2 Geographically Weighted Regression Algorithm

This algorithm is a form of spatial analysis that uses statistical data by assigning them as neighbor values for a set location and calculating the relation between the two, trying to minimize the value to reach an ideal point [2]. Generally, GWRfollows the following equation,

$$y_i u = \beta_{0i} u + \beta_{1i} u x_{1i} + \beta_{2i} u x_{2i} + ... + \beta_{mi} u x_{mi} \tag{4.4}$$

Here, y is the dependent variable. At location u, the dependent variable is regressed on m number of independent variables which are denoted by x. As discussed earlier, this algorithm uses statistical data for creating a relationship between the location and its surrounding. Here, β is that relationship.

The main idea of Geographically Weighted Regression is to employ the idea of local models, where they are considered subsets of the main model and one single focal point is fixed. From the earlier equation, the dependent variable y is calculated for each of these local or subset of models, and regression is performed accordingly.

4.1.3 Lasso Regression Algorithm

This regression algorithm is a linear regression that works on the idea of shrinkage. LASSO is an acronym for Least Absolute Shrinkage and Selection Operator [9]. Lasso regression has widely been used for some time now as an algorithm to optimize statistical models using a penalty system. This model prevents overfitting of data by assigning a penalty for such data, and trying to minimize this penalty in each iteration for the whole dataset. This mode of penalty minimization is also known as L1 regularization. The value of the penalty is the absolute value of the magnitude of coefficients. The general formula of the lasso regression is as follows,

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \gamma \sum_{j=1}^{p} |\beta_j| \tag{4.5}$$

Here, γ is a tuning parameter, which determines the penalty and its effect on the result. A higher value of γ refers to more coefficients being eliminated and an increase in bias as the total value starts to become more dependent on the remaining variables. On the other hand, a decrease in the value of γ tells us that the variance is increasing as the total value starts to depend on more and more variables. When this value of γ reaches 0, no parameters are needed to be deleted and the estimate is correct.

4.1.4 Elastic Net Regression Algorithm

Just like lasso regression, Elastic Net Regression algorithm uses a penalty system to perform regression and eliminate over-fitting. Unlike lasso regression however, Elastic Net Regression has two penalty terms rather than one [6]. The basic formula for this regression is,

$$min_{\beta 0*\beta}(\frac{1}{2N}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2 + \gamma P_a(\beta)) \tag{4.6}$$

Here, x is the independent variable, y is the dependent variable, β_0 is the intercept and value of $Pa\beta$ is calculated using the following formula,

$$P_a(\beta) = \frac{(1-a)}{2}||\beta||_2^2 = \sum_{j=1}^{p}(\frac{(1-a}{2}\beta_j^2 + \alpha|\beta_j|) \tag{4.7}$$

Here, α is a parameter that determines the mixing of the two penalty terms. The two penalty terms are $||\beta||$ 1 which is called the L1 norm of β and $||\beta||2$ which is called the L2 norm of β. This algorithm works by trying to minimize the value of $P_a(\beta)$ by assigning the values of L1 and L2 norms and getting a minimized value of β intercept

4.1.5 Bi-LSTM Algorithm

Bi-LSTM refers to Bidirectional Long Short Term Memory. This is a type of neural network that processes data in both directions, thus called bidirectional [12]. In traditional neural networks, the relation between nodes is short-term, and the relation between nodes of further distances are not observed. To overcome this shortcoming, the idea of LSTM was introduced. This observes and stores the relation between distant nodes to generate a better neural network.

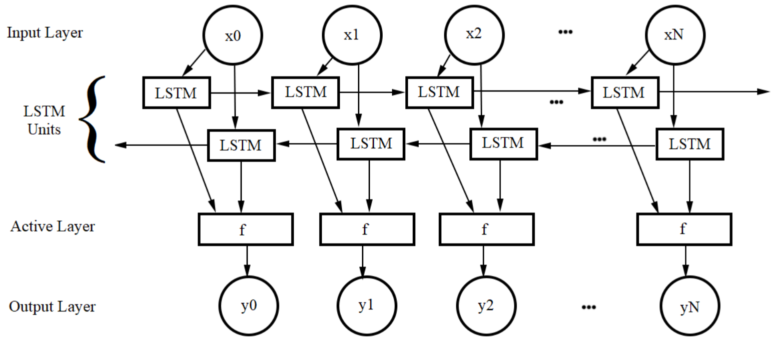

Figure 4.2: Bi-LSTM Structure

There are 4 core steps of the Bi-LSTM algorithm:

Input: The dataset is given to the model. Each data point is then separated into different nodes and represented as a vector.

Embedding: The imputed values are converted into dense vector representations, which form a neural network for the algorithm to work on.

Bi-LSTM: The LSTM operation is run in both forward and backward directions simultaneously with different sets of parameters, comprising two separate layers.

Output: The two layers from before are merged to form a single output which contains the hidden findings from the two layers to give a more accurate result.

4.1.6 DiffSynth Tab

This model is a form of DiffSynth model appropriate for tabular data. As we are working with such tabular data in our research, we have used the DiffSynth Tab model instead of DiffSynth model, which usually is used on audio or video data. However, the models are similar in their working principle. From paper[29], we can learn the following about the model:

- Probabilistic models known as diffusion models use the reversal of a diffusion process to learn how to produce data. Starting with noise, they iteratively improve outputs to produce structured data such as photos, audio, or, in this example, tablature. The diffusion model is taught to translate latent noise into intelligible tablature sequences in DiffSynth Tab.
- For stringed instruments, tabs provide an organized representation of the music that includes information on tempo, frets, and string numbers. The model represents tabs as sequences using a special encoding approach, making them appropriate for neural networks.
- To efficiently handle sequential data, the architecture usually integrates recurrent layers or transformers with the diffusion process. It may involve extra processes, such as attention layers, for comprehending temporal and harmonic links in music.

4.1.7 Feed Forward Neural Network

From the paper[1], we have learnt about the basic structure and working principle of Feedforward Neural Network or FNN.

The input layer takes in the data, such as the dataset's features. Hidden layers transform the input data. Nodes (neurons) with corresponding activation functions make up these levels. Depending on the job, the output layer generates the final prediction or choice (e.g., classification probabilities or regression results).

The data flows in one direction unlike Bi-LSTM. Some of the activation functions used in FNN are:

$$\text{Sigmoid: } f(x) = \frac{1}{1+e^{-x}} \tag{4.8}$$

$$\text{ReLU (Rectified Linear Unit): } f(x) = \max(0, x) \tag{4.9}$$

$$\text{Tanh: } f(x) = \tanh(x) \tag{4.10}$$

The strength of every neuronal connection is determined by its corresponding weight. To increase the flexibility of neurons' decision-making process, bias words are included.

Forward Propagation uses the network to send input data to calculate the output. Backward Propagation adjusts weights and biases by calculating the gradient of the loss function with respect to each parameter. Optimization procedure reduces the loss function (such as Mean Squared Error 'MSE') using techniques such as Adam, RMSprop, or Gradient Descent.

4.1.8 Tabular Score Based Model

As we are working with tabular data in our research, we have used normal score based models that work on tabular data, in this case, Tabular Score Based Model[23]. The basic idea of this model is the following:

Structure: Made up of columns (features or variables) and rows (examples, instances). Numerical, category, ordinal, and text data can all be considered feature types. Models in which every instance in the dataset is assigned a score. The score is frequently used as a basis for decision-making tasks such as ranking or prioritization. Relative importance, utility values, or probabilities can all be represented by the score.

Decision trees, random forests, and gradient boosted trees (such as XGBoost, LightGBM, and CatBoost) are examples of tree-based models. They are ideal for tabular data because they can handle a variety of feature kinds and successfully capture feature interactions.

Linear Models: To model the relationships between variables, use either logistic regression or linear regression with feature engineering.

Deep Learning Models: Neural networks specifically tailored to tabular data, like TabNet or DeepFM, that use deep learning.

Probabilistic Models: Depending on the probability of a certain output, scores are assigned.

4.1.9 Gaussian Perturbation Model

A mathematical framework on the Gaussian Distribution[4], Gaussian Perturbation Model is often used in diffusion modeling.

The Gaussian distribution, sometimes referred to as the normal distribution, is a probability distribution with a bell shape that is distinguished by its standard deviation (σ) and mean (μ). In the context of models, perturbation is defined as small, random changes put into a system, usually to account for noise or simulate unpredictability.

The model uses random variables selected from a Gaussian distribution to represent the perturbations. Depending on the application, the model implies that these perturbations are either multiplicative or additive:

$$\text{Additive:} y = f(x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{4.11}$$

$$\text{Multiplicative:} y = f(x)(1 + \epsilon), \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{4.12}$$

In systems where deviations are anticipated to be regularly distributed around a central value, this technique is used to model noise. Frequent in situations where Gaussian-distributed noise results from the application of the Central Limit Theorem. In generative modeling, Gaussian perturbations are applied to data iteratively before being reversed to recreate structured outputs.

To explore the solution space more efficiently, the following perturbation formula is used:

$$y = f(x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \Sigma) \tag{4.13}$$

$$J(x) = J_0(x) + \mathcal{N}(0, \sigma^2) \tag{4.14}$$

4.1.10 TabDDPM

Unlike most other generative diffusion models, TabDDPM is suitable for tabular data[24]. The key points we have learnt about this model are:

Diffusion models are probabilistic generative models that reverse a slow noise process to learn to create data.

The model effectively learns the data distribution during training by predicting how to denoise a given noisy sample. The data is transformed into random noise by gradually adding noise. After that, the model gradually reassembles the original data from the noise.

Adding Gaussian noise (for continuous features) or a discrete noising technique (for categorical features) gradually taints tabular data. The data is converted into a series of noisy versions. Using a neural network, usually parameterized by a transformer-like architecture or a UNet, the model learns to reverse the noise process. Synthetic tabular data is reconstructed from noise in this step.

Then using a loss function, usually MSE or cross-entropy, the noising and denoising process is fine tuned gradually.

4.1.11 DiffFlow

The name DiffFlow is a combination of the word ‘Diffusion’ and ‘Flows’. The DiffFlow model is a generative diffusion model that incorporates normalizing flows into the diffusion process for better performance[28].

Normalizing Flows: A class of generative models that employ invertible transformations to convert a simple probability distribution (such as a Gaussian) to a complicated data distribution. The transformation can translate data to and from a latent space without loss. Flows are extremely efficient in computing likelihoods, making them ideal for likelihood-based modeling.

DiffFlow Hybridization: Combines diffusion models' characteristics (robustness and sample quality) with the expressiveness of normalized flows. Introduces a flow-based transformation into the diffusion model's forward and reverse processes to boost modeling capacity.

The model is trained to minimize a loss function that combines diffusion loss and log-likelihood estimation (normalizing flows). This dual purpose guarantees both high-quality generation and a tractable probabilistic framework.

4.1.12 Effective Distance Model

A mathematical framework for analyzing distance between nodes that correlate with other factors in the dataset[16]. This model is often used for epidemic research as it can effectively identify the disease spread from simple tabular data.

Beyond physical or geographic distance, the model considers the impact of network topology or dynamic processes on the network. Reflects how "reachable" one node is from another, taking into account traffic flow, infection rates, and communication strength.

The effective distance is frequently calculated using weights assigned to network edges, which might reflect probabilities, costs, or flows. Weights are typically inversely connected to the ease of interaction; for example, a lower weight may suggest a stronger connection or a higher likelihood of interaction. For nodes i and j, where weight is Wij:

$$\text{Effective Distance}(i, j) \propto -\log(w_{ij}) \tag{4.15}$$

Effective distance is frequently calculated using shortest-path algorithms (such as Dijkstra's or Bellman-Ford), but it replaces actual distances with effective distances obtained from network features.

Used to simulate the spread of infectious diseases by calculating the effective distance between locations depending on travel habits.

4.2 TOPSIS & VIKOR

Techniques like VIKOR (VIseKriterijumska Optimizacija I Kompromisno Resenje) and TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) are used to determine the weights for the factors: Total_admitted, Death, and Released. Using TOPSIS and VIKOR, all the rows in the dataset has been ranked.

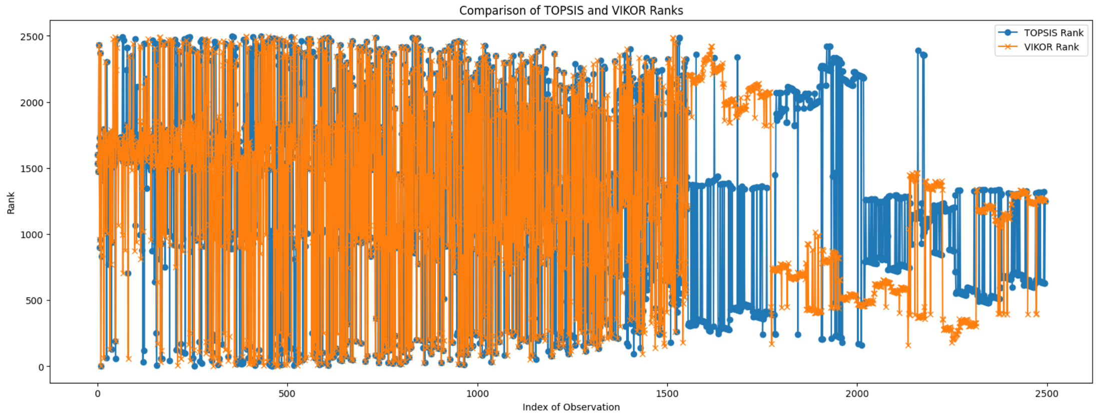

Figure 4.3: Comparison of TOPSIS and VIKOR Ranks

Then,using the following algorithm, entropy based weights have been calculated from the ranked dataset:

Algorithm 1 Entropy-based Weight Calculation

```
1: **Input:** Data ranked using TOPSIS & VIKOR
2: Normalize the data: $\texttt{normalized\_data} = \texttt{data} \ / \ \texttt{max(data)}$
3: Calculate probabilistic distribution (p):
       $$\texttt{p} = \texttt{normalized\_data}/\texttt{sum(normalized\_data)}$$
4: Compute entropy:
       $$\texttt{entropy} = -\sum \left(\texttt{p} \times \log(\texttt{p} + 10^{-9})\right) / \log(\texttt{len(data)})$$
5: Calculate weights:
       $$\texttt{weights} = \frac{1 - \texttt{entropy}}{\sum(1 - \texttt{entropy})}$$
6: **Output:** Weight values for target features
```

Here in Step-2, the data is normalized so that the value of each data point remains between 0 and 1.

For calculating entropy, we need to multiply the value of p with its log value. A small value (1e - 9) is added to avoid log of zero. Also the negative value of sum is considered as entropy is defined to be a positive quantity.

Chapter 5

Performance Evaluation

5.1 Performance Metrics

Mean Squared Error (MSE): MSE measures the average squared difference between the actual and predicted values. A lower MSE indicates that the model's predictions are closer to the true values, making it a key metric for regression models.

R-squared ($\mathbf{R^2}$): R^2, or the coefficient of determination, evaluates how well the independent variables explain the variance in the dependent variable. It ranges from 0 to 1, where a higher value signifies a better fit of the model to the data.

Accuracy: Accuracy is a performance metric commonly used for classification models. It represents the proportion of correctly classified instances out of the total instances, providing an overall measure of the model's effectiveness.

In this study, we utilized MSE, R^2, and Accuracy to evaluate and compare multiple models. By analyzing these performance metrics, we identified the best-performing model, which was then used to refine the weighting process for our Effective Distance Model in predicting dengue hotspots.

Metrics like Precision, Recall, and F1-Score were deemed irrelevant for assessing model performance because this is a regression problem. The main application of these metrics is in classification tasks, where classifying instances into different classes is the aim. Regression issues, on the other hand, need the prediction of continuous values, therefore metrics like Mean Squared Error (MSE) and R-squared are better suited for evaluating how well the model predicts numerical results. While R-squared aids in quantifying the percentage of variance explained by the model, MSE offers a clear measure of prediction error. We make sure that our evaluation criteria are in line with the nature of the issue and the kind of predictions we are generating by concentrating on these regression-specific indicators.

5.1.1 Performance Analysis

Models	Total Admitted MSE	Total Admitted R^2	Total Admitted Accuracy	Death MSE	Death R^2	Death Accuracy	Released MSE	Released R^2	Released Accuracy
OLS		0.024			0.010			0.023	
GWR	1169964.97	-0.0517		75.8655	-0.4832		1158455.77	-0.0544	
LASSO Regression	528456.59	0.0234	0.0	26.8069	0.0030	0.298	487628.92	0.0207	0.004
Elastic-Net Regression	525580.23	0.0287	0.004	26.7896	0.0036	0.302	484623.06	0.0267	0.0
Bi-LSTM	563317.65	-0.0410	0.29	27.0329	-0.0054	0.572	519012.70	-0.0424	0.138
DiffSynthTab	24658.19	0.9544	0.9544	7.05	0.7379	0.7379	38041.22	0.9236	0.9236
Feedforward Neural Network (FNN)	4004.30	0.9926	0.9926	19.99	0.2566	0.2566	1796.30	0.8842	0.8842
Tabular Score-Based Model	57141.90	0.8944	0.8944	75331.33	0.8487	0.8487	53.56	0.9922	0.9922
Gaussian Perturbation Neural Network	5290.03	0.9902	0.9902	25.33	0.0579	0.0579	1679.99	0.8917	0.8917
TabDDPM	69481.61	0.8716	0.8716	72.03	-1.6790	-1.6790	94251.20	0.8107	0.8107
DiffFlow	127632.30	0.7641	0.7641	9.86	0.6332	0.6332	139395.25	0.7200	0.7200

Table 5.1: Comparison of Models Based on Various Metrics

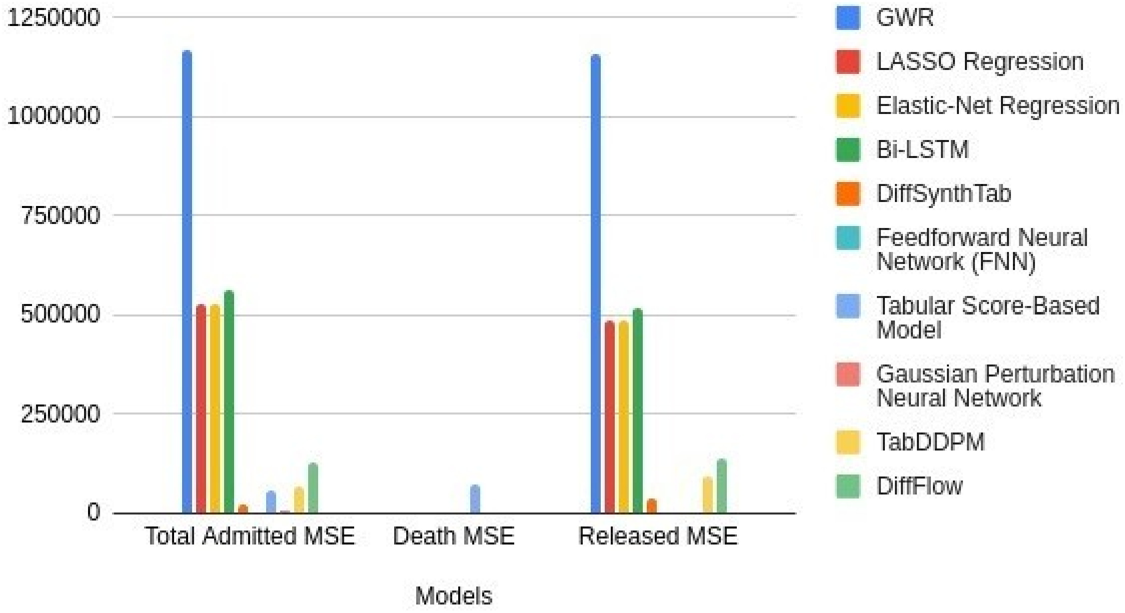

Figure 5.1: Total Admitted MSE, Death MSE, Released MSE

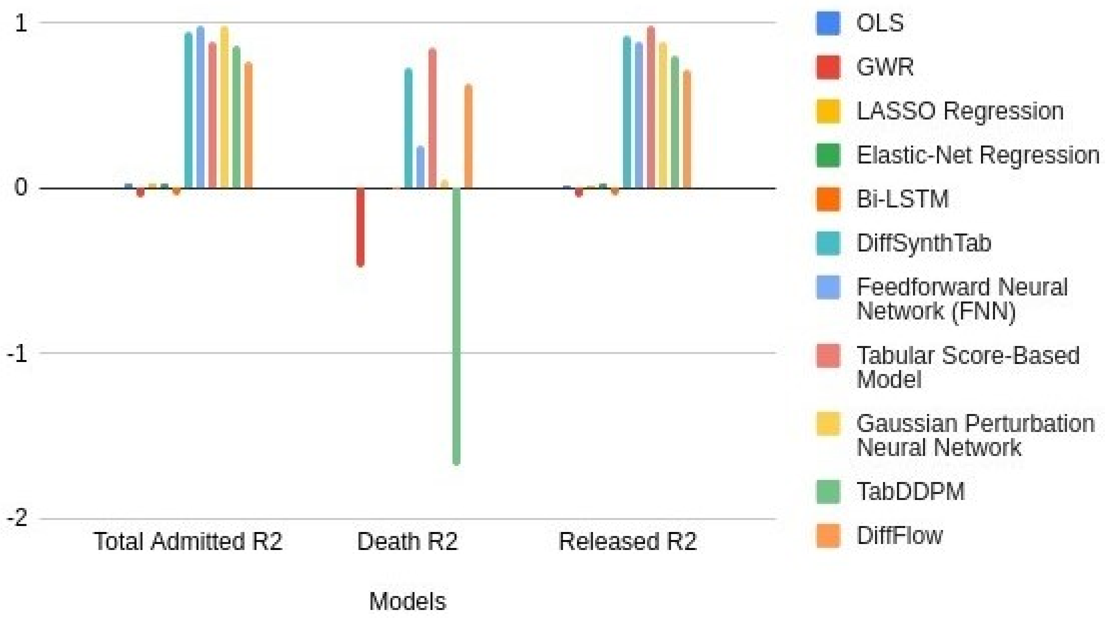

Figure 5.2: Total Admitted R2, Death R2, Released R2

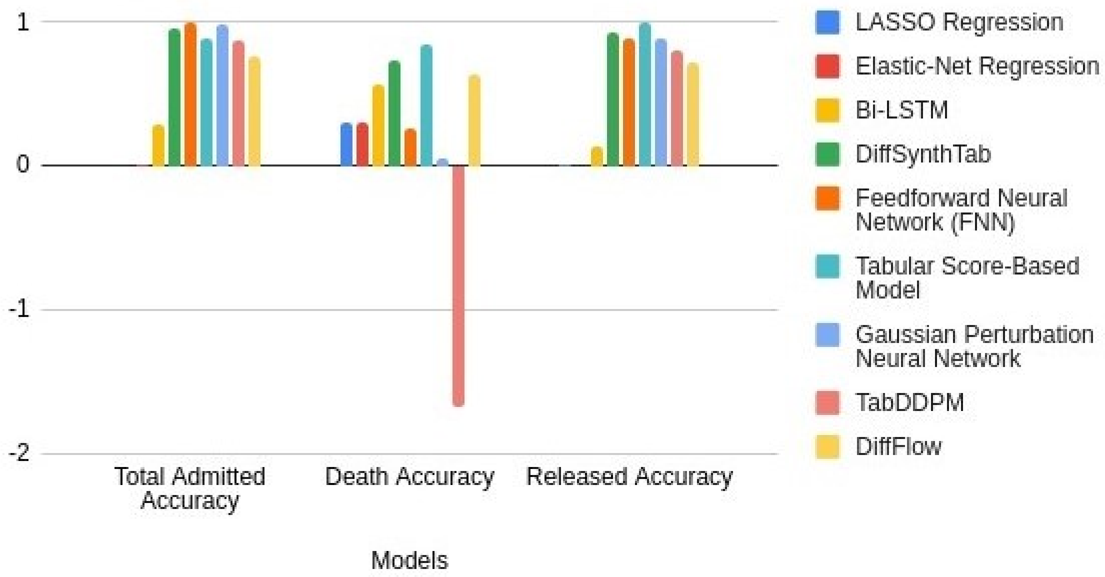

Figure 5.3: Total Admitted Accuracy, Death Accuracy, Released Accuracy

5.1.2 Comparative Study

This study evaluates various machine learning and diffusion models for predicting dengue cases, focusing on three key target variables: Total Admitted, Death, and Released. The models were assessed using three primary performance metrics: Mean Squared Error (MSE), R-squared (R^2), and Accuracy. The objective was to determine the best-performing model and utilize its derived weights to enhance the Effective Distance Model for identifying dengue hotspots.

Model Performance Analysis

1. **Traditional Regression Models**

   - **Ordinary Least Squares (OLS):** OLS showed relatively poor performance across all target variables, with high MSE values and low R^2 scores.

   - **Geographically Weighted Regression (GWR):** GWR demonstrated significantly high MSE values, particularly for the Total Admitted and Released categories. Negative R^2 values indicate that the model failed to fit the data adequately.

   - **LASSO and Elastic-Net Regression:** Both models showed slight improvements over OLS, with marginally better R^2 scores. However, their overall predictive capabilities remained weak, particularly for the Death and Released categories.

2. **Deep Learning-Based Models**

   - **Bi-LSTM (Bidirectional Long Short-Term Memory):** Despite showing some promise with an accuracy of 0.29 for Total Admitted, it struggled with negative R^2 values for the other target variables, indicating poor generalization.

   - **Feedforward Neural Network (FNN):** FNN performed significantly better, achieving high R^2 and accuracy scores across all variables. Its performance on Total Admitted was particularly strong ($R^2 = 0.9926$, Accuracy = 0.9926).

   - **Gaussian Perturbation Neural Network:** The Gaussian Perturbation Neural Network demonstrated moderate performance, showing some improvements over traditional regression models. However, its overall effectiveness was not as strong as the best-performing models, making it a less optimal choice for dengue prediction.

3. **Diffusion-Based Models**

   - **DiffSynthTab:** This model outperformed most traditional approaches, with high R^2 scores (0.9544 for Total Admitted, 0.7379 for Death, and 0.9236 for Released). Its accuracy values were also significantly better than regression models.

   - **Tabular Score-Based Model:** This was the best-performing model, achieving an R^2 of 0.8944 for Total Admitted, 0.8487 for Death, and 0.8107 for Released. It demonstrated superior predictive capability compared to other models.

   - **TabDDPM & DiffFlow:** These models performed well but had inconsistencies, particularly in the Death category, where TabDDPM had a negative R^2 score (-1.6790). DiffFlow showed a moderate performance but lagged behind the Tabular Score-Based Model in accuracy and overall reliability.

Based on the comparative analysis, the **Tabular Score-Based Model** was selected as the best-performing model. The weights for Total Admitted, Death, and Released were computed based on its performance. These weights were then combined with the TOPSIS and VIKOR rankings to determine final weights, which were subsequently used in the Effective Distance Model for hotspot detection and prediction. This approach ensures a more reliable and data-driven method for identifying dengue-prone areas and assisting authorities in disease control and awareness efforts.

5.2 Dengue Hotspot

5.2.1 Weight Calculation

As we have seen from the formula for calculating effective distance, we need weight values for the calculation. In our research, we have obtained this weight value from two different sources and combined them using a weighted average method.

Firstly, we have extracted the weight values for our required features from the best performing machine learning model, tabular score based model. Next, we have used TOPSIS and VIKOR to rank the data and calculate entropy based weight as seen before in chapter 4.4.

Secondly, these two derived weights for each target features have been combined using the following weighted average formula, where x is the weight value from TOPSIS and VIKOR and y is the weight value from the best performing machine learning model:

$$\text{Weighted\_average} = \frac{(x \times \text{weight } 1) + (y \times \text{weight } 2)}{\text{weight } 1 + \text{weight } 2} \tag{5.1}$$

Using this formula, the weight values have been calculated for all three features: "Total_admitted", "Released", and "Death".

5.2.2 Effective Distance Calculation

For the calculation of the effective distance between districts, we calculate a risk factor between two districts using the following formula[16], which is later used to calculate the effective distance between district A and district B:

$$\begin{aligned} \text{Risk\_factor} = & \left[\frac{\text{Total\_admitted\_A} + \text{Total\_admitted\_B}}{2} \times \text{weight}\right] \\ & - \left[\frac{\text{Released\_A} + \text{Released\_B}}{2} \times \text{weight}\right] \\ & + \left[\frac{\text{Death\_A} + \text{Death\_B}}{2} \times \text{weight}\right] \end{aligned} \tag{5.2}$$

$$\text{Effective\_distance} = \frac{\text{distance}}{1 + \text{Risk\_factor}} \tag{5.3}$$

5.2.3 Dengue Hotspot Analysis

The effective distance graph is created, and the nodes are connected where the effective distance is lower than the threshold value 0.1. After testing with multiple threshold values, we concluded that the value 0.1 is the most accurate for hotspot creation, as it creates edges between nodes that are actually facing higher-than-average dengue cases per month.

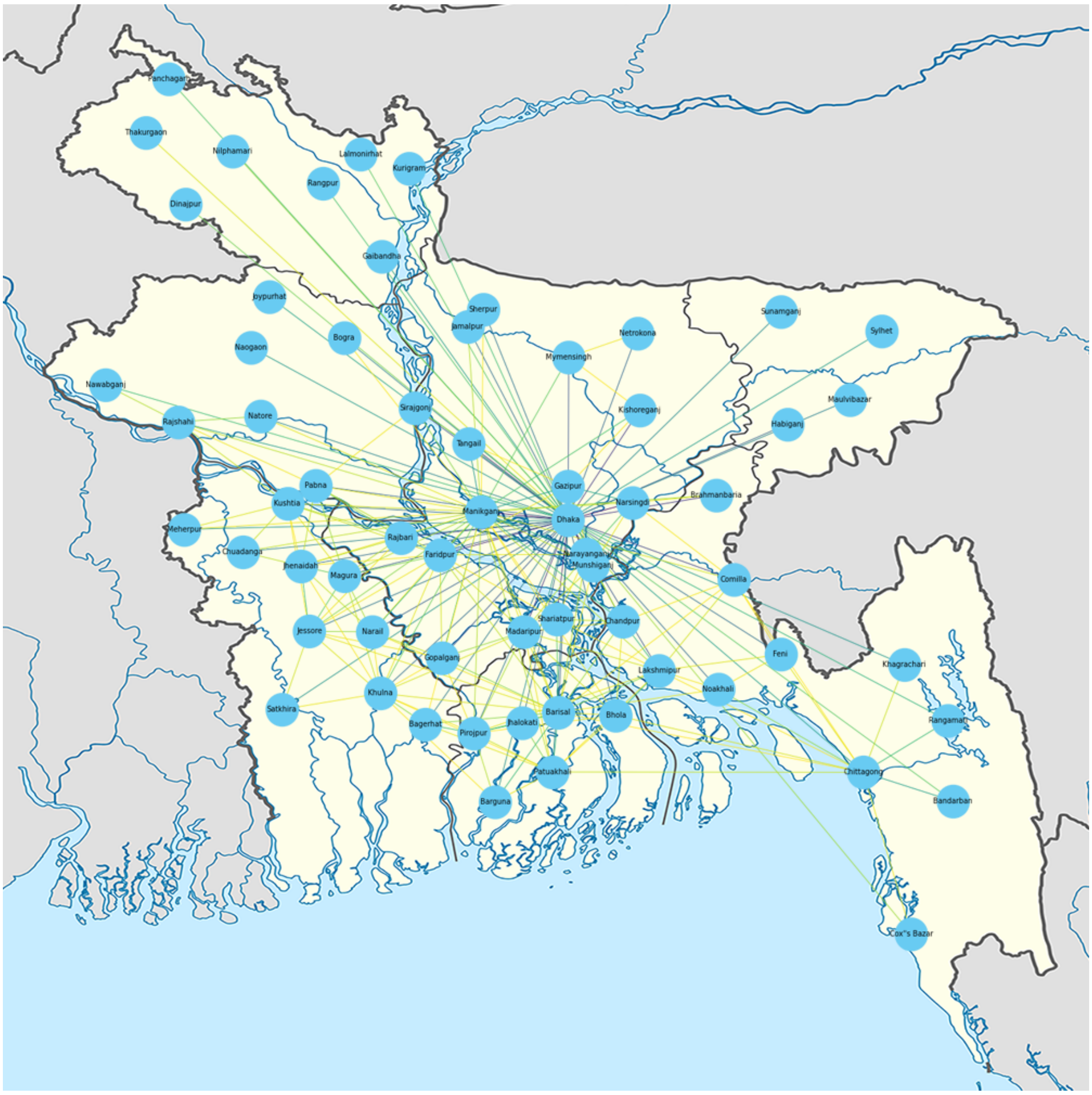

Figure 5.4: Effective Distance Network for Dengue Diffusion Patterns

Then, for detecting hotspots, we have used the top 4% of the districts with the highest patient count as this also correlates with the observed data and marks districts that are actually experiencing an influx of dengue patients.

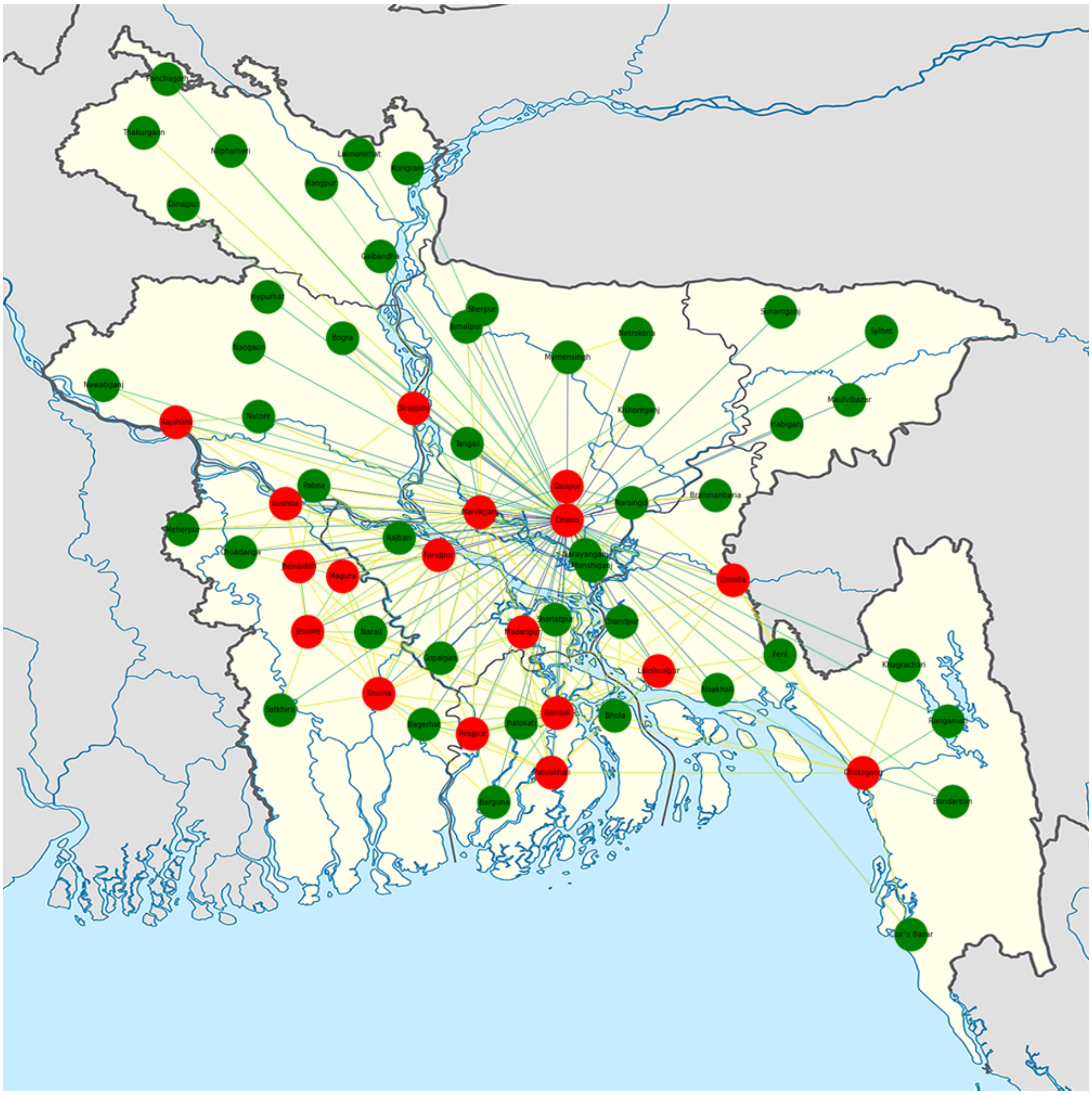

Figure 5.5: Effective Distance Network with High-Risk Districts Highlighted in Red

5.2.4 Effective Distance Verification

Conducting a validity study on the ED model means that the final map of high-risk dengue districts is safe from errors. In other words, we should compare them with other machine learning models such as Logistic Regression and Random Forest to check the matching level of the prediction of ED. Thus, this verification process aids in establishing that ED is indeed a reliable procedure, and any attempts made to manipulate the findings would provide reduced efficiency and quality for the method. It also offers assurance in the deployment of the model as inform of decision making tool for the public health.

Effective Distance with Machine learning Model (Random Forest)

In order to check The Effective Distance (ED) model, the classification of high-risk dengue districts by both the ED model and Random Forest (RF) model were compared. For the analyses, the epidemiological parameters Total Admitted, Released, and Death were selected as the principle performance features for developing the RF model for high-risk districts. Risk analysis or how districts were classified By using Total admitted cases in the ED model, and using the 96 percentile to classify them as high risk, the high risk option left out most areas with low risk, thus avoiding most of the false positives. The RF model was applied using 8%0-20% cross-validation data that helps the model to learn from historical dengue data and predict the risk level on unseen data. Its accuracy was computed as:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \times 100 \tag{5.4}$$

resulting in an accuracy of over 80% which makes it a valid risk prediction model.In order to compare the ED and RF classifications, a matching coefficient was derived as a percentage of the total.

$$\text{ED-ML Agreement} = \frac{\sum(\text{ED Classification} = \text{ML Prediction})}{\text{Total Districts}} \times 100 \tag{5.5}$$

The last check of the correlation between the ED model and the RF model was 71%, meaning that the ED model's high-risk categories were quite similar to the ML-based predictions.

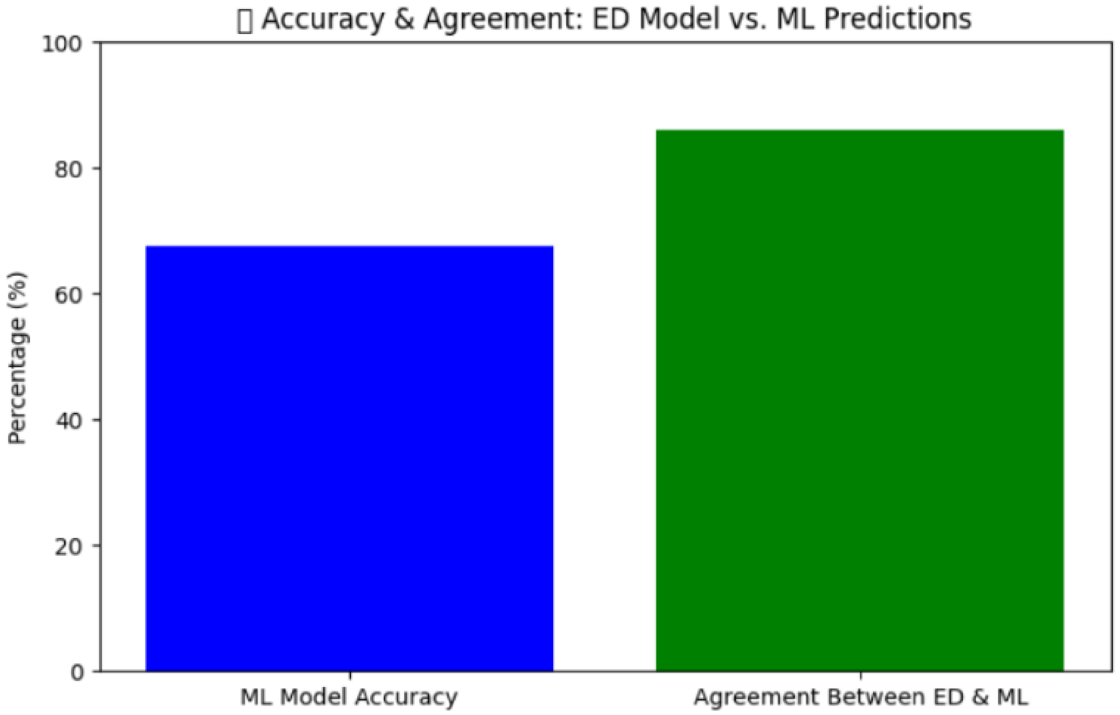

Figure 5.6: Effective Distance with ML Model (Random Forest)

1. If RF Model Accuracy $\geq$ 80%, it is a reliable risk prediction model.
2. If ED-ML Agreement $\geq$ 70%, the ED model is considered accurate and does not require modification.
3. If ED-ML Agreement $<$ 50%, the ED model would need weight tuning and improvements.

Since the ED-ML agreement was above 70%, the ED model is confirmed to be robust with statistical legitimacy for determining districts with the high risk of dengue outbreaks. Heatmap was used to represent the level of classification similarity of the ED and RF models; bar chart represents the percentage of accuracy achieved in RF model as well as the percentage of agreement circumferenced by ED and ML. Pursuant to this, no modifications are warranted for the ED model and the model can be confidently used for risk categorization of the epidemiological outbreaks of dengue.

Effective Distance with Machine learning Model (Logistic Regression)

The ED model was compared to the LogReg model and both the models achieved 71 percent accuracy to predict high-risk dengue districts. The accuracy was computed as:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \times 100 \tag{5.6}$$

The agreement between ED and LogReg was 71.04%, calculated as:

$$\text{ED-ML Agreement} = \frac{\sum(\text{ED Classification} = \text{ML Prediction})}{\text{Total Districts}} \times 100 \tag{5.7}$$

Logistic Regression and the Effectiveness Display Model had almost the same accuracy level merely slightly over 71%. Logistic Regression shows that ED has 71.04% of agreement with the ED model, and thus ED's high-risk predictions are comparable to ML results. If in any case the agreement was below 50% then it would demand weight tuning as well as improvements in the ED model. Since the final agreement result is 71.04%, the ED model can be concluded as an effective method in identifying the districts with high risk of dengue fever without further alteration.

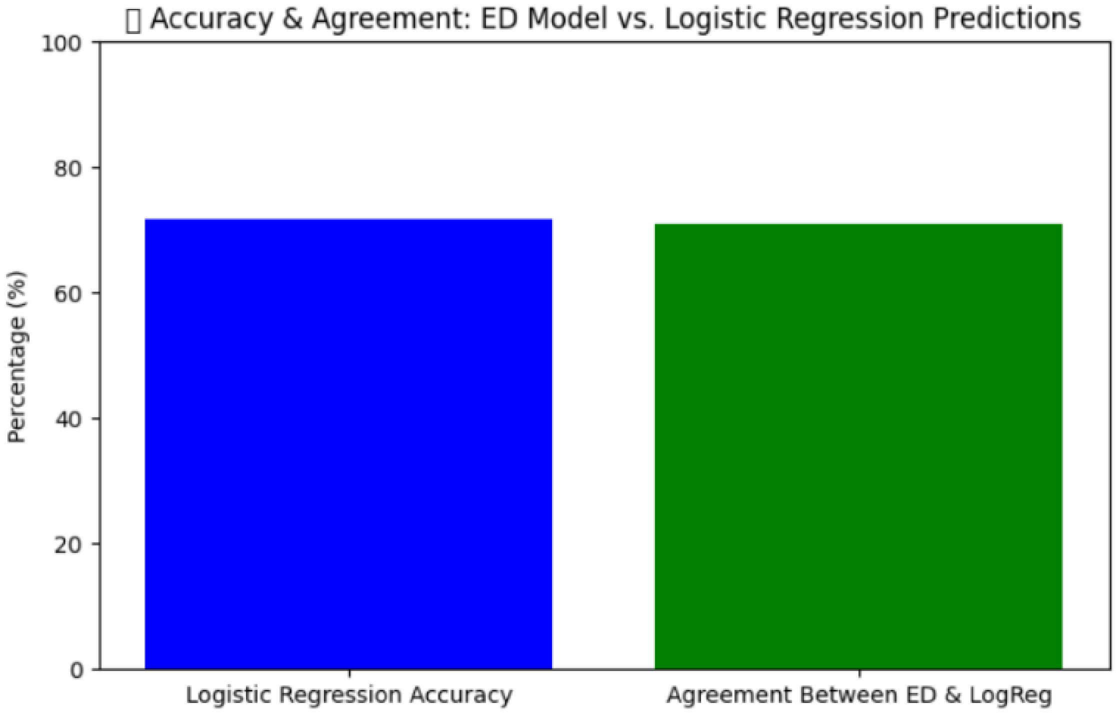

Figure 5.7: Effective Distance with ML Model (Logistic Regression)

Logistic Regression and ED model both achieved accuracy slightly above 71ED model showed 71.04Below 50With 71.04

- Logistic Regression and ED model both achieved accuracy slightly above 71%.
- ED model showed 71.04% agreement with Logistic Regression, confirming its effectiveness.
- Below 50% agreement would require model adjustments.
- With 71.04% agreement, ED is validated as an effective method for identifying high-risk dengue districts.

Chapter 6

Conclusion

This research encountered numerous challenges, underscoring the complexity of studying dengue diffusion patterns in Bangladesh. Key difficulties included data quality and availability, as the dataset lacked comprehensive temporal and geographic coverage, making it difficult to conduct a detailed analysis of dengue dynamics. During the COVID-19 pandemic, there were major data gaps and difficulties in gathering accurate patient and weather data. Even with the use of sophisticated imputation methods like Kriging and KNN, producing consistent and realistic data was still challenging.

The interdependencies and non-linear correlations between environmental variables including temperature, humidity, and rainfall made feature selection difficult. It took a lot of feature engineering to determine which elements had the biggest effects, which made the workflow more complex. Robust hyperparameter tuning was required for sophisticated models such as Gaussian Perturbation Neural Networks and Tabular Score-Based Models, and there were additional challenges in striking a balance between model complexity and interpretability for public health applications.

Weak temporal dependencies in the dataset caused problems for sequential models like Bi-LSTM, requiring creative data representation strategies. Additionally, data sparsity limited insights into the socio-environmental dynamics of illness diffusion by making it difficult to integrate socio-environmental elements like population density and healthcare access. To properly investigate dengue diffusion, these difficulties underscore the urgent need for enhanced data collection methods and reliable modeling approaches.

6.1 Limitations and Future Works

This research highlights the potential of machine learning and geospatial techniques in analyzing dengue diffusion patterns in Bangladesh. The primary objectives were to optimize dengue diffusion using machine learning models and identify dengue clusters to aid public health planning and community awareness. The research indicates that the integration of sophisticated models with geographical analysis can increase public health planning, facilitate targeted interventions, and improve detecting spreading patterns. However, there are still problems that need to be fixed with feature selection, data preparation, and model tuning. The environmental, geographic and climate aspects that influence dengue diffusion need to be more precisely included in models.

Future research should concentrate on enhancing data collecting by integrating larger datasets that include comprehensive socioeconomic, geographic and temporal characteristics, like population movement and healthcare access. By simulating actual diffusion settings, generative approaches like DiffSynthTab and TabDDPM, as well as sequential models like Bi-LSTM, can be explored and optimized to further improve analysis. By incorporating these improved models into public health systems, early warning systems can be made available, allowing for prompt treatments and effective use of resources. In order to enable comprehensive epidemic management and aid in worldwide disease control efforts, this framework can be expanded to include other vector-borne diseases in addition to real-time data and automated decision-support tools.

Bibliography

[1] G. Bebis and M. Georgiopoulos, “Feed-forward neural networks,” *IEEE Potentials*, vol. 13, no. 4, pp. 27–31, 1994. DOI: 10.1109/45.329294.

[2] A. Fotheringham, M. Charlton, and C. Brunsdon, “Geographically weighted regression: A natural evolution of the expansion method for spatial data analysis,” *Environment and Planning A*, vol. 30, pp. 1905–1927, Feb. 1998. DOI: 10.1068/a301905.

[3] C. Dismuke and R. Lindrooth, “Ordinary least squares,” *Methods and Designs for Outcomes Research*, Jan. 2006.

[4] G. Papandreou and A. L. Yuille, “Gaussian sampling by local perturbations,” in *Advances in Neural Information Processing Systems*, J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., vol. 23, Curran Associates, Inc., 2010. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2010/file/d09bf41544a3365a46c9077ebb5e35c3-Paper.pdf.

[5] P. Jeefoo, N. K. Tripathi, and M. Souris, “Spatio-temporal diffusion pattern and hotspot detection of dengue in chachoengsao province, thailand,” *International Journal of Environmental Research and Public Health*, vol. 8, no. 1, pp. 51–74, 2011, ISSN: 1660-4601. DOI: 10.3390/ijerph8010051. [Online]. Available: https://www.mdpi.com/1660-4601/8/1/51.

[6] C. Hans, “Elastic net regression modeling with the orthant normal prior,” *Journal of the American Statistical Association*, vol. 106, p. 1383, Jan. 2012. DOI: https://doi.org/10.1198/jasa.2011.tm09241.

[7] J. E. M. P. Pessanha, W. T. Caiaffa, M. C. d. M. Almeida, S. T. Brandão, and F. A. Proietti, “Diffusion pattern and hotspot detection of dengue in belo horizonte, minas gerais, brazil,” *Journal of Tropical Medicine*, vol. 2012, no. 1, p. 760 951, 2012. DOI: https://doi.org/10.1155/2012/760951. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2012/760951. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2012/760951.

[8] S. Atique, T.-C. Chan, C.-C. Chen, *et al.*, “Investigating spatio-temporal distribution and diffusion patterns of the dengue outbreak in swat, pakistan,” *Journal of Infection and Public Health*, vol. 11, no. 4, pp. 550–557, 2018, ISSN: 1876-0341. DOI: https://doi.org/10.1016/j.jiph.2017.12.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1876034117302915.

[9] J. A. C. J Ranstam, “Lasso regression,” *British Journal of Surgery*, vol. 105, p. 1348, Sep. 2018. DOI: https://doi.org/10.1002/bjs.10895.

[10] M. I. Naiyar Iqbal, "Machine learning for dengue outbreak prediction: A performance evaluation of different prominent classifiers," *Informatica, An international journal of computing and informatics*, 2019. DOI: https://doi.org/10.31449/inf.v43i3.1548.

[11] R. Kamesh and N. Sivakumar, "Design and implementation of realtime detection of dengue using machine learning," *Journal of Advanced Research in Dynamical and Control Systems*, vol. 12, Jun. 2020. DOI: 10.5373/JARDCS/V12SP5/20201851.

[12] M. Massaoudi, S. S. Refaat, H. Abu-Rub, I. Chihi, and F. Oueslati, "Pls-cnn-bilstm: An end-to-end algorithm-based savitzky-golay smoothing and evolution strategy for load forecasting," *Energies*, vol. 13, p. 29, Oct. 2020. DOI: 10.3390/en13205464.

[13] D. Sarma, S. Hossain, T. Mittra, A. Bhuiya, I. Saha, and R. Chakma, "Dengue prediction using machine learning algorithms," Dec. 2020, pp. 1–6. DOI: 10.1109/R10-HTC49770.2020.9357035.

[14] S. Chattopadhyay, A. Chattopadhyay, and E. Aifantis, "Predicting case fatality of dengue epidemic: Statistical machine learning towards a virtual doctor," *Journal of Nanotechnology in Diagnosis and Treatment*, vol. 7, pp. 10–24, Oct. 2021. DOI: 10.12974/2311-8792.2021.07.2. [Online]. Available: http://savvysciencepublisher.com/jms/index.php/jndt/article/view/751.

[15] M. Islam, S. Khushbu, A. Azad Rabby, and T. Bhuiyan, "A study on dengue fever in bangladesh: Predicting the probability of dengue infection with external behavior with machine learning," English, Proceedings - 5th International Conference on Intelligent Computing and Control Systems, ICICCS 2021 ; Conference date: 06-05-2021, May 2021, pp. 1717–1721. DOI: 10.1109/iciccs51141.2021.9432288.

[16] Q. Shang, Y. Deng, and K. H. Cheong, "Identifying influential nodes in complex networks: Effective distance gravity model," *Information Sciences*, vol. 577, pp. 162–179, 2021, ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2021.01.053. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025521000918.

[17] M. R. Al Nasar, I. Nasir, T. Mohamed, N. S. Elmitwally, M. M. Al-Sakhnini, and T. Asgher, "Detection of dengue disease empowered with fused machine learning," in *2022 International Conference on Cyber Resilience (ICCR)*, 2022, pp. 01–10. DOI: 10.1109/ICCR56254.2022.9996009.

[18] S. K. Dey, M. M. Rahman, A. Howlader, *et al.*, "Prediction of dengue incidents using hospitalized patients, metrological and socio-economic data in bangladesh: A machine learning approach," *PLOS ONE*, vol. 17, no. 7, pp. 1–17, Jul. 2022. DOI: 10.1371/journal.pone.0270933. [Online]. Available: https://doi.org/10.1371/journal.pone.0270933.

[19] D. M. A. Khan, J. Akter, I. Ahammad, S. Ejaz, and T. Khan, "Dengue outbreaks prediction in bangladesh perspective using distinct multilayer perceptron nn and decision tree," *Health Information Science and Systems*, vol. 10, Nov. 2022. DOI: 10.1007/s13755-022-00202-x.

[20] M. T. Sarwar and M. Al Mamun, “Prediction of dengue using machine learning algorithms: Case study dhaka,” in *2022 4th International Conference on Electrical, Computer And Telecommunication Engineering (ICECTE)*, 2022, pp. 1–6. DOI: 10.1109/ICECTE57896.2022.10114535.

[21] G. Gupta, S. Khan, V. Guleria, *et al.*, “Ddpm: A dengue disease prediction and diagnosis model using sentiment analysis and machine learning algorithms,” *Diagnostics*, vol. 13, no. 6, 2023, ISSN: 2075-4418. DOI: 10.3390/diagnostics13061093. [Online]. Available: https://www.mdpi.com/2075-4418/13/6/1093.

[22] Z. Hussain, I. Khan, and M. Arsalan, “Machine learning approaches for dengue prediction: A review of algorithms and applications,” vol. 78, pp. 15–36, Jun. 2023.

[23] J. Kim, C. Lee, and N. Park, *Stasy: Score-based tabular data synthesis*, 2023. arXiv: 2210.04018 `[cs.LG]`. [Online]. Available: https://arxiv.org/abs/2210.04018.

[24] A. Kotelnikov, D. Baranchuk, I. Rubachev, and A. Babenko, “TabDDPM: Modelling tabular data with diffusion models,” in *Proceedings of the 40th International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., ser. Proceedings of Machine Learning Research, vol. 202, PMLR, 23–29 Jul 2023, pp. 17 564–17 579. [Online]. Available: https://proceedings.mlr.press/v202/kotelnikov23a.html.

[25] M. A. Majeed, H. Z. M. Shafri, Z. Zulkafli, and A. Wayayok, “A deep learning approach for dengue fever prediction in malaysia using lstm with spatial attention,” *International Journal of Environmental Research and Public Health*, vol. 20, no. 5, 2023, ISSN: 1660-4601. DOI: 10.3390/ijerph20054130. [Online]. Available: https://www.mdpi.com/1660-4601/20/5/4130.

[26] S. Raheja, S. Kasturia, X. Cheng, and M. Kumar, “Machine learning-based diffusion model for prediction of coronavirus-19 outbreak,” *Neural Comput and Applic*, vol. 35, 2023, ISSN: 13755–13774. DOI: https://doi.org/10.1007/s00521-021-06376-x. [Online]. Available: https://link.springer.com/article/10.1007/s00521-021-06376-x.

[27] N. Siddique, M. Arefin, M. Kaiser, and A. Kayes, “Applied intelligence for industry 4.0,” 2023. DOI: https://doi.org/10.1201/9781003256083. [Online]. Available: https://www.taylorfrancis.com/books/edit/10.1201/9781003256083/applied-intelligence-industry-4-0-nazmul-siddique-mohammad-shamsul-arefin-shamim-kaiser-asm-kayes.

[28] J. Zhang, H. Shi, J. Yu, E. Xie, and Z. Li, *Diffflow: A unified sde framework for score-based diffusion models and generative adversarial networks*, 2023. arXiv: 2307.02159 `[stat.ML]`. [Online]. Available: https://arxiv.org/abs/2307.02159.

[29] Z. Duan, L. You, C. Wang, *et al.*, “Diffsynth: Latent in-iteration deflickering for realistic video synthesis,” in *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track*, A. Bifet, T. Krilavičius, I. Miliou, and S. Nowaczyk, Eds., Cham: Springer Nature Switzerland, 2024, pp. 332–347, ISBN: 978-3-031-70381-2.