Understanding the Dynamics of Dengue in Bangladesh: EDA, Climate Correlation, and Predictive Modeling

Sabrina Masum Meem
Department of CSC
Independent University, Bangladesh
Dhaka, Bangladesh
1720694@iub.edu.bd

Mohammed Tahmid Hossain
Department of CSC
Independent University, Bangladesh
Dhaka, Bangladesh
1820228@iub.edu.bd

Jannat Khair Chowdhury
Department of CSE
Independent University, Bangladesh
Dhaka, Bangladesh
1830105@iub.edu.bd

Md Saef Ullah Miah
Department of Computer Science, FST
American International University-Bangladesh (AIUB)
Dhaka, Bangladesh
md.saefullah@gmail.com

Md Fahad Monir
Department of CSE
Independent University, Bangladesh
Dhaka, Bangladesh
fahad.monir@iub.edu.bd

***Abstract*—Dengue, a mosquito-borne viral infection, poses a significant threat, especially in warm, tropical climate countries like Bangladesh, India, Thailand, Malaysia, Laos, etc. This study is solely focused on the dengue data of Bangladesh as it explores the historical dengue data spanning 23 years (2000 to 2022) for EDA purposes, with a focus on 9 years (2014-2022) divisional data for model performance analysis. Additionally, climate data was collected for the same period to examine the potential correlation between dengue cases and climate factors. Machine learning (ML) and Deep learning (DL) models, including Random Forest Regression (RFR), Long Short-Term Memory (LSTM), and LSTM with Artificial Neural Networks (ANN), were implemented and validated against ground truth data. The results reveal notable differences in performance between ML and DL models when handling imbalanced datasets with outliers, with RFR outperforming LSTM when compared to the ground truth data. The study uncovers significant correlations between dengue cases and climate factors like humidity, temperature, and precipitation. The insights gained from this research have practical implications for dengue prevention and control efforts in Bangladesh and beyond, paving the way for more effective strategies and interventions.**

***Index Terms*—Dengue, Climate Change, Time series analysis, Random Forest Regression (RFR), Long Short-Term Memory (LSTM), Deep Learning**

I. Introduction

Dengue is a mosquito-borne viral infection that poses a significant threat, especially in warm, tropical climates. Dengue fever is a flu-like illness characterized by high fever, severe headache, joint and muscle pain, rash, and sometimes life-threatening complications. One of the most alarming aspects of dengue is its ability to cause severe forms of the disease, such as dengue hemorrhagic fever (DHF) and dengue shock syndrome (DSS), which can lead to organ damage, bleeding, and even death if not promptly treated [4]. Each year, approximately 390 million people experience fever, and 96 million require hospitalization due to dengue fever [9]. Dhaka, the country's capital with a population of over 16 million, reported the highest number of dengue cases from 2012 to 2019 [2]. The monsoon season in Bangladesh occurs from May to September, providing favorable conditions for the proliferation of the Aedes mosquito and increasing its density in Bangladesh. Additionally, a common practice of storing water in household containers like buckets and pots is a common practice in Bangladesh, contributing significantly to the breeding of Aedes mosquitoes. This creates ideal breeding grounds for Aedes mosquitoes [1]. During the COVID-19 lockdown in 2020, it was found, there was an approximately 98% drop in dengue cases in Bangladesh. Following the lifting of the lockdown in 2021, there was a 95% increase in the number of dengue cases.

Our research focuses on a comprehensive analysis of Dengue trends in Bangladesh using a dataset spanning 22 years. The primary objective is to investigate the correlation between the country's climate patterns and the occurrence of Dengue cases. Additionally, we aim to assess the effectiveness and reliability of ML and DL models in predicting Dengue outbreaks based on our dataset. The contributions of our study can be summarized as follows:

- The original dataset was in the Bangla language, requiring us to perform a manual translation to create an English dataset. This translation process demanded meticulous attention to detail and accuracy to ensure the validity of the translated information. By undertaking this effort, we gained valuable insights into the dataset.
- We systematically organized the daily district-wise dataset into eight divisions and calculated the division-wise monthly averages. This division-based approach

resulted in the creation of eight distinct datasets. Such segregation allowed us to examine and analyze the dengue trends within each specific division, enabling more focused and granular insights.

- Having completed extensive data preprocessing, we have successfully compiled a comprehensive divisional dataset for Bangladesh, encompassing a period of 9 years. Recognizing the importance of collaborative research and knowledge sharing, we have made the dataset openly accessible to fellow researchers. By providing access to this valuable resource, we aim to foster scientific advancements and encourage further investigations in the field of Dengue research within the country.
- To further deepen our analysis, we employed web scraping techniques using Beautiful Soup to acquire the historical climate data for each division. This enabled us to establish a comparison and identify potential correlations between dengue cases and climate patterns across the country. This additional layer of analysis offered valuable insights into the relationship between dengue occurrences and climate variations.

II. Related Works

A significant number of studies across the world explore the use of ML techniques in predicting and analyzing dengue fever outbreaks in various regions. They shed light on the contributing factors to dengue transmission, including climate change, urbanization, and inadequate vector control. A case study was conducted in China [5] to develop a dengue forecast model using ML techniques. The study utilizes various factors, such as climate variables and historical dengue data, to create an accurate prediction model. The Support Vector Regression (SVR) algorithm aims to achieve accurate predictions and robust generalization even in complex and noisy data. The model detected the peak of the 2014 outbreak and tracked dengue dynamics in different cities. The authors in [13] found that most individuals afflicted with dengue were between the ages of 15 and 49, and 80% of cases are within urban communities, with Selangor being a densely populated and urban state in Malaysia. Using ML algorithms, the results indicated that rainfall played a role in creating mosquito larva habitats, with increased and decreased rainfall contributing to dengue outbreaks. Furthermore, Sheng-Wen Huang et al. [7] focused on predicting the risk of dengue severity using ML techniques, including Linear Regression, Support Vector Machine, Random Forest, Gradient Boosting Method, and Artificial Neural Network (ANN). The results revealed that utilization of the ANN method yielded a model with good prediction performance.

Several studies have been conducted on Bangladesh as dengue is a significant public health challenge due to its high incidence rates and mortality.

In [9], authors collected climate, water storage, and hygiene data from diverse regions of Bangladesh. By incorporating external behaviors, they demonstrated Machine Learning's capability for accurate dengue predictions, with Naive Bayes, Decision Tree, and Random Forest models achieving the highest accuracy. Another study [3] was conducted in Bangladesh and used ML to predict dengue fever, where the Decision Tree achieved the highest accuracy. The authors in [14] used meteorological and socio-economic data to predict dengue outbreaks by integrating various data sources. The study found that the Random Forest algorithm performed the best among the tested algorithms. The authors in [2] found that due to the COVID-19 pandemic, it had been reported a significant drop in dengue fever cases in 2020 compared to 2019, which saw an unprecedented spike with over 100,000 cases. As with various stereotypes of dengue, the study [10] found that the risk of severe dengue increases when multiple stereotypes are present, posing a significant threat to future outbreaks.

During the 5-year study period, [11] noted the highest percentage of severe dengue cases in 2019, primarily linked to the prevalence of DENV3. Additionally, [8] discovered a correlation between meteorological data and dengue cases, with the monsoon season in Bangladesh showing the highest number of positive cases due to mosquito-friendly rainy weather.

This literature review emphasizes the unique challenges faced by multiple countries including Bangladesh due to their geographical and climatic conditions, making them particularly vulnerable to dengue outbreaks. Our study exclusively focuses on understanding the significance of comprehending climate associations with Dengue over an extensive period of time and effectively managing outliers within the existing body of research. Moreover, we aim to provide a comprehensive overview of the performance of ML and DL techniques for handling outliers as well as to address the significant public health challenge posed by dengue in Bangladesh.

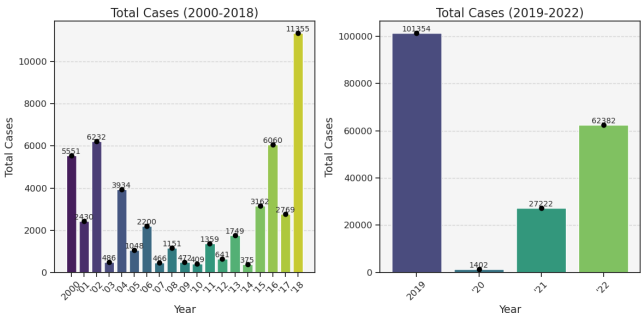


Fig. 1: Reported Dengue Cases from 2000-2022

III. Data Description

A. Data Source and Study Area

We collected data from the Department of Health and Family Welfare of the Government of the People's Republic of Bangladesh to conduct our research. The study period spans from the year 2000 to 2023. The dataset includes information from 63 districts in Bangladesh, covering all

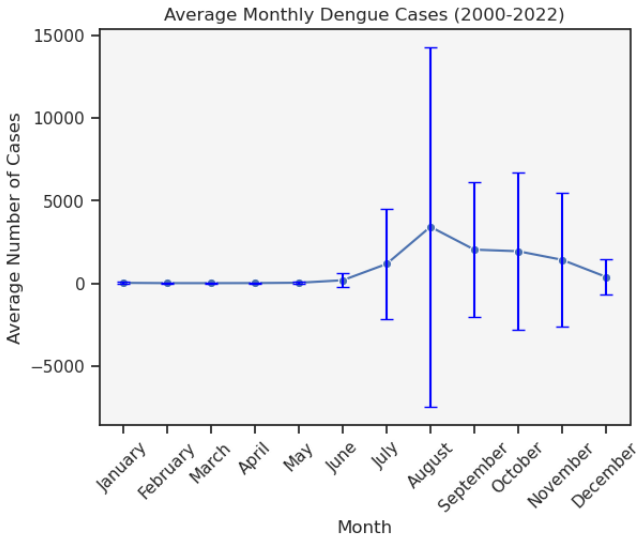


Fig. 2: Trend of Average Monthly Dengue Cases (2000-2022)

eight divisions: Dhaka, Mymensingh, Chittagong, Khulna, Rajshahi, Rangpur, Barisal, and Sylhet.

B. Data Pre-processing

The dataset is sourced from the Ministry of Health & Family Welfare of the People's Republic of Bangladesh. Initially, the dataset was in the Bangla language and covered monthly records of dengue cases and deaths in Bangladesh from 2000 to 2022. These records were utilized solely for exploratory data analysis (EDA) in our study. Additionally, the dataset included district-wise dengue records from January 1st, 2014, to April 30th, 2023, with daily records for each 24-hour period across all districts. This district-wise dataset provided information on the regions where dengue cases were found, along with the total number of deaths. For model training, we focused on a divisional dataset derived from district-wise data.

Dhaka, Bangladesh's capital city, had more detailed case information, including data from public and private hospitals. The dataset comprised various columns: symptoms of dengue within 24 hours, such as Dengue Fever (DF), Dengue Shock Syndrome (DSS), and Dengue Hemorrhagic Fever (DHF); the number of individuals affected by each symptom; the total number of possible cases; the actual number of patients admitted to hospitals; the total death cases; the number of currently admitted patients; and the total number of patients who recovered and were discharged.

The entire dataset comprising Bangla language content necessitated translation into English. Our initial endeavor to use the dataset in Python encountered technical challenges. Consequently, we explored the possibility of automatic translation employing diverse fonts within MS Word, Excel documents, and Google Translate. Regrettably, it was discerned that the employed font (SutonnyMJ) for preserving the original data possessed encoding anomalies, impeding automated translation. Therefore, we resorted to manual translation of the data.

To prepare the divisional dengue dataset, monthly divisional data from January 1st, 2014, to April 30th, 2023, was extracted from the daily district-wise data. The selected columns for this study included the total number of monthly dengue cases and the total number of monthly deaths caused by dengue for each year. This resulted in a divisional dataset encompassing the eight divisions of the country.

For comparison analysis with climate data, we extracted daily historical climate data from 2014 to 2023 using web scraping techniques with Beautiful Soup. The climate data consisted of the Date, Region, and daily values for temperature, humidity, and precipitation for each of the eight divisions of the country. After pre-processing the data by checking and dropping null values, the daily climate data was aggregated into monthly average values.

Finally, the divisional dengue and climate data were merged based on the Date and Region columns, creating a single dataset that contained information on both dengue and climate. As the region column contained categorical values, label encoding was applied to convert these values into numerical ones suitable for our study. The final dataset consisted of 855 records and included seven columns: Date, Region Encoded, Average Temperature, Average Precipitation, Average Humidity, Total Cases, and Total Deaths. The dataset was sorted chronologically, facilitating time series analysis.

IV. Methodology

Our study was divided into two primary phases: data preparation and pre-processing for the exploratory data analysis (EDA) in the first phase and the implementation and validation of ML and DL models in the second phase. The overview of the methodology is shown in Fig. 4.

Bangladesh has faced numerous dengue fever outbreaks over the years, with varying numbers of reported cases and severity of outbreaks. In recent times, dengue has posed a significant threat to the country, with a high number of cases. We visualized the data using a bar plot to analyze the dengue situation in Bangladesh over the years. The dataset used for visualization included total monthly cases from 2000 to 2022 reported in the country, providing a comprehensive overview of dengue trends over the past 22 years (2000-2022) shown in Fig. 1. We split the figures into two plots, one containing 2000-2018 dengue cases and the other containing 2019-2022 cases, as there were significant differences in the number of total cases during the recent years.

It is to be mentioned that our dataset experienced an unprecedented surge in reported dengue cases in 2019, which was likely caused by the introduction of a new serotype (DENV-3), an unusually wet pre-monsoon season, and assumed resistance to insecticides [16]. The total cases exceeded 100,000, deviating significantly from the normal data distribution. It is crucial to handle these outliers appropriately, as they can negatively impact the accuracy of

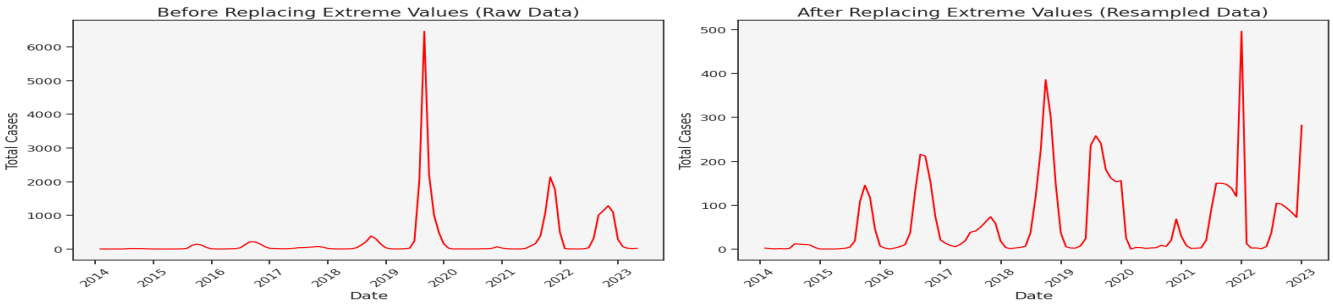


Fig. 3: Raw data VS Resampled data

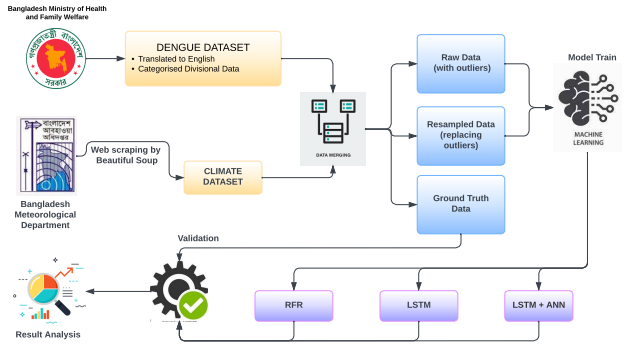


Fig. 4: Workflow of our project

training and prediction models by deviating from the original data patterns.

We created a resampled version of the original dataset to mitigate the influence of outliers. The rationale behind this approach was to replace the total reported case count for each month in 2019 with the corresponding monthly averages from the previous outlier-free years (2014-2018, 2020). We used these resampled values to preserve the overall data patterns observed in the non-outlier years while accounting for the exceptional spike in 2019. Additionally, we observed a sudden drop in dengue cases in the early months of 2020, likely due to the nationwide lockdown to prevent the spread of the Covid-19 pandemic [6]. However, towards the end of 2020, the reported dengue cases began to increase at a growth rate approximately 15% higher than the previous years before 2019. Consequently, 2021 and 2022 were also identified as outliers in our dataset, similar to 2019. To mitigate the immediate impact of these outliers, we utilized averages from a comparable time period, further minimizing distortion caused by outliers and enhancing the underlying pattern in our dataset. This step aimed to reduce the sensitivity of DL models to outliers and improve the accuracy of their predictions which is shown in Fig. 3.

For the purpose of validation, the ground-truth dataset was created with the raw dengue and climate data for the period from January 2023 to April 2023. To summarize, in this study, we employed three distinct datasets: *raw data*, which contains the original data; *resampled data*, which includes the outlier-free data and Ground Truth Data, which encompasses the validated and authentic data used for comparison and verification. Subsequently, RFR, LSTM and LSTM-ANN models were trained on both the raw and resampled data.

V. Result Analysis & Discussion

This section is divided into two parts: Part V-A Analysis of the dengue situation and Part V-B Prediction Model analysis.

A. Analysis of the dengue situation

We categorized our data into eight divisions of Bangladesh. According to the Bangladesh Bureau of Statistics (BBS), the Dhaka division has the highest population density of 2,147 people/km^2, and the Barishal division has the lowest population density of 688 people/km^2 [15]. From the dataset covering 2014-2022, we observed a total of 1,55,809 dengue cases in Dhaka and 10,913 cases in Barishal, indicating a potential association between population density and dengue outbreaks. Our primary focus was on analyzing the impact of three climate factors (Precipitation, Humidity, and Temperature) on dengue transmission and identifying outliers in the dataset.

We conducted a correlation analysis generating a heatmap as shown in Fig-5, which revealed humidity amplifies dengue transmission, followed by temperature and precipitation. The analysis of Fig-2 indicates that the period from July to early November, with August being the peak month, exhibits the highest incidence of Dengue cases. Our examination of the climate data reveals that these months correspond to the most humid period in the country, characterized by significantly greater rainfall compared to other months. These findings highlight a robust correlation between humidity and the spread of Dengue cases. Total Death had the highest correlation with Total Cases but as it is a natural phenomenon, this study did not consider this correlation.

B. Prediction Model Analysis

For this study, our aim was to predict the total possible cases for the first four months of 2023 (Jan-April) by implementing ML and DL models. And secondly, to compare and validate whether DL or ML models work better on our

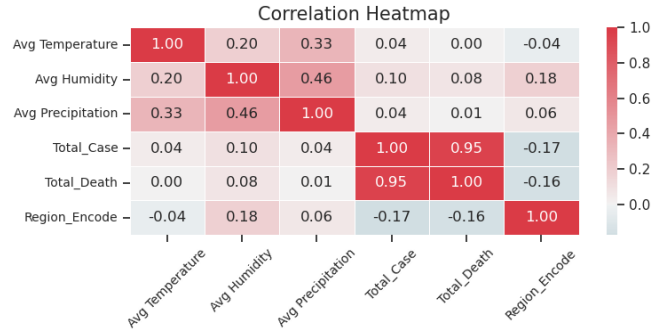


Fig. 5: Correlation Heatmap

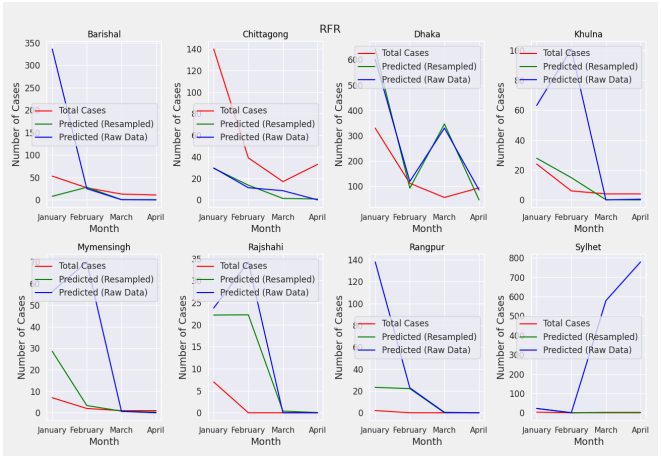
datasets. We choose RFR, LSTM, and LSTM combined with the ANN approach to train both our raw and resampled data. The implemented models are trained on both the raw and resampled dataset. The values for January - April 2023 were predicted to validate with the ground truth dataset. The results shown in Fig- 6 states that RFR outperformed LSTM AND LSTM-ANN implementations on both the raw and resampled dataset.

As our datasets are imbalanced and contain a significant amount of zeros and outliers, these potentially disrupt the sequence's regular patterns, making it an obstacle for our model to discern genuine trends from noise.

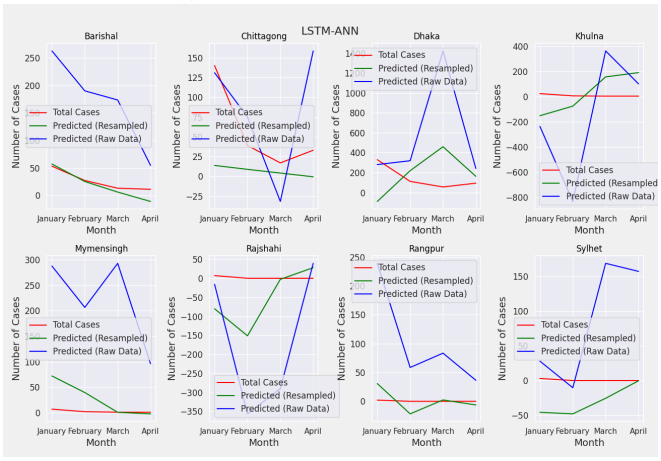
The poor performance of the LSTM model could be attributed to its difficulty in capturing temporal dependencies and patterns in data. DL models tend to perform better for larger training datasets because they have a more significant number of parameters allowing them to learn a broader range of patterns [12]. In our case, the raw and resampled training dataset has limited outlier instances, so the model may not learn to handle these instances properly during the training phase. LSTM has some limitations while capturing complex nonlinear relationships and handling irregular patterns in data. To leverage the strengths of the two models, we combined ANN with LSTM. By incorporating two models: ANN's ability to capture nonlinear dependencies and benefit from LSTM's sequential analysis, the combination allowed the models to capture the complex interaction better, ultimately leading to improved prediction accuracy than LSTM only. Still, it did not yield significant improvements compared to RFR. As Fig. 6 shows, both DL models predicted negative values, whereas RFR predicted non-negative values. The full-size images of all the plots are available here: Project_Dengue

In our DL model, we implemented the nonlinear activation function 'ReLu', which outputs the input value if it is positive and zero otherwise. So when the outlier with a significant positive value was present in the dataset, the ReLU function output the value as it overpowered the contribution of other data points. As the effect of these outliers is magnified after applying this function may be resulted to have significant influence leading to a distorted prediction.

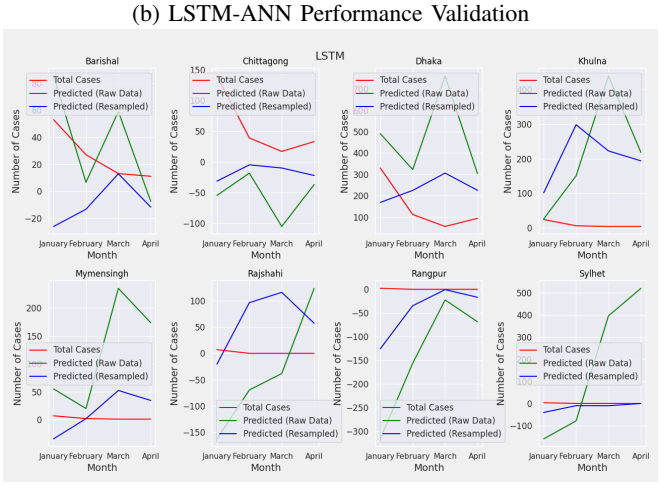
As RFR, a traditional ML model is based on ensemble



(a) RFR Performance Validation



(b) LSTM-ANN Performance Validation



(c) LSTM Performance Validation

Fig. 6: Model Comparison

learning made of multiple decision trees. Decision trees are the building block of RFR. Each decision tree learns from a different subset of data and can handle nonlinear relationships by recognizing complex patterns among features [17]. Outliers are often isolated in their leaf nodes rather than significantly impacting overall prediction. While making

predictions, the wisdom of all decision trees is combined, taking into account the average majority vote of prediction from each different leaf, which dissolves the influence of outliers as a whole. This robustness leverages the RFR model to handle outliers more effectively than DL models. However, the limitation of this model for this complex dataset of dengue transmission is that the model might have assumed a linear relationship already exists between our input features and target variable, which in our case, is not true. And while the nonlinear relationship is high, this model may struggle to recognize dengue patterns and outliers.

VI. Conclusion

Overall, this study provides insights into the dengue situation in Bangladesh and demonstrates the potential of using ML and DL techniques to address outliers. We analyzed the climate and dengue data across the eight divisions from which an increasing dengue trend was found all over Bangladesh. Additionally, our heatmap demonstrates the correlation between climate variables (humidity, temperature, precipitation) and dengue cases, considering other influencing factors. We developed a prediction model incorporating divisional dengue and historical weather data. We trained a traditional machine learning model (RFR), deep learning models (LSTM), and a combined deep learning model (LSTM with ANN) using Temperature, humidity, and precipitation as independent variables, which were identified as crucial factors in dengue transmission. The model results exhibited promising performance, with RFR demonstrating superior outlier handling compared to other models. In future research, we aim to integrate statistical modeling techniques to enhance outlier handling within the data. By understanding the factors that contribute to dengue outbreaks and implementing predictive models, in the future, more measures can be taken to control and reduce the impact of this disease. This knowledge can assist public health authorities in implementing effective strategies for preventing and controlling this mosquito-borne viral infection.

References

[1] Dengue – Bangladesh. [Accessed 10-Jun-2023].

[2] Khokon Kanti Bhowmik, Jannatul Ferdous, Prodip Kumar Baral, and Mohammad Safiqul Islam. Recent outbreak of dengue in bangladesh: A threat to public health. *Health Science Reports*, 6(4):e1210, 2023.

[3] Samrat Kumar Dey, Md Mahbubur Rahman, Arpita Howlader, Umme Raihan Siddiqi, Khandaker Mohammad Mohi Uddin, Rownak Borhan, and Elias Ur Rahman. Prediction of dengue incidents using hospitalized patients, metrological and socio-economic data in bangladesh: A machine learning approach. *PLoS One*, 17(7):e0270933, 2022.

[4] Saif Mahmud Khan Dourjoy, Abu Mohammed Golam Rabbani Rafi, Zerin Nasrin Tumpa, and Mohd Saifuzzaman. A comparative study on prediction of dengue fever using machine learning algorithm. In *Advances in Distributed Computing and Machine Learning: Proceedings of ICADCML 2020*, pages 501–510. Springer, 2021.

[5] Pi Guo, Tao Liu, Qin Zhang, Li Wang, Jianpeng Xiao, Qingying Zhang, Ganfeng Luo, Zhihao Li, Jianfeng He, Yonghui Zhang, et al. Developing a dengue forecast model using machine learning: A case study in china. *PLoS neglected tropical diseases*, 11(10):e0005973, 2017.

[6] Mohammed Tahmid Hossain, Afra Hossain, Sabrina Masum Meem, Md Fahad Monir, Md Saef Ullah Miah, and Talha Bin Sarwar. Impact of covid-19 lockdowns on air quality in bangladesh: Analysis and aqi forecasting with support vector regression. In *2023 4th International Conference for Emerging Technology (INCET)*, pages 1–6, 2023.

[7] Sheng-Wen Huang, Huey-Pin Tsai, Su-Jhen Hung, Wen-Chien Ko, and Jen-Ren Wang. Assessing the risk of dengue severity using demographic information and laboratory test results with machine learning. *PLoS neglected tropical diseases*, 14(12):e0008960, 2020.

[8] Md Aminul Islam, Mohammad Nayeem Hasan, Ananda Tiwari, Md Abdul Wahid Raju, Fateha Jannat, Sarawut Sangkham, Mahaad Issa Shammas, Prabhakar Sharma, Prosun Bhattacharya, and Manish Kumar. Correlation of dengue and meteorological factors in bangladesh: A public health concern. *International Journal of Environmental Research and Public Health*, 20(6):5152, 2023.

[9] Md Sanzidul Islam, Sharun Akter Khushbu, Akm Shahariar Azad Rabby, and Touhid Bhuiyan. A study on dengue fever in bangladesh: Predicting the probability of dengue infection with external behavior with machine learning. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1717–1721. IEEE, 2021.

[10] Mohammad Enamul Hoque Kayesh, Ibrahim Khalil, Michinori Kohara, and Kyoko Tsukiyama-Kohara. Increasing dengue burden and severe dengue risk in bangladesh: An overview. *Tropical Medicine and Infectious Disease*, 8(1):32, 2023.

[11] Rummana Rahim, Abu Hasan, Juthamas Phadungsombat, Nazmul Hasan, Nikhat Ara, Suma Mita Biswas, Emi E Nakayama, Mizanur Rahman, and Tatsuo Shioda. Genetic analysis of dengue virus in severe and non-severe cases in dhaka, bangladesh, in 2018–2022. *Viruses*, 15(5):1144, 2023.

[12] M Saef Ullah Miah and Junaida Sulaiman. Material named entity recognition (mner) for knowledge-driven materials using deep learning approach. In *Proceedings of the Fourth International Conference on Trends in Computational and Cognitive Engineering: TCCE 2022*, pages 199–208. Springer, 2023.

[13] Nurul Azam Mohd Salim, Yap Bee Wah, Caitlynn Reeves, Madison Smith, Wan Fairos Wan Yaacob, Rose Nani Mudin, Rahmat Dapari, Nik Nur Fatin Fatihah Sapri, and Ubydul Haque. Prediction of dengue outbreak in selangor malaysia using machine learning techniques. *Scientific reports*, 11(1):939, 2021.

[14] Dhiman Sarma, Sohrab Hossain, Tanni Mittra, Md Abdul Motaleb Bhuiya, Ishita Saha, and Ravina Chakma. Dengue prediction using machine learning algorithms. In *2020 IEEE 8th R10 humanitarian technology conference (R10-HTC)*, pages 1–6. IEEE, 2020.

[15] Bangladesh Statistics. Bangladesh bureau of statistics (bbs). *Statistics and Informatics Division, Ministry of Planning www. bbs. gov. bd.(Accessed 9 November 2019)*, 2018.

[16] Snigdha Rahman Titir, Shyamal Kumar Paul, Salma Ahmed, Nazia Haque, Syeda Anjuman Nasreen, Khondoker Shoaib Hossain, Fahim Uddin Ahmad, Sultana Shabnam Nila, Jobyda Khanam, Neaz Nowsher, et al. Nationwide distribution of dengue virus type 3 (denv-3) genotype i and emergence of denv-3 genotype iii during the 2019 outbreak in bangladesh. *Tropical medicine and infectious disease*, 6(2):58, 2021.

[17] M. Saef Ullah Miah, Junaida Sulaiman, Kamal Z. Zamli, Shah Samiur Rashid, and Ahmed Jalal Khan Chowdhury. Predicting carboxymethyl cellulase assay (cmcase) production using artificial neural network and explicit feature selection approach. In *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, pages 1–6, 2023.