Prediction of Dengue Cases in Bangladesh using Explainable Machine Learning Approach

Conference Paper · June 2024

DOI: 10.1109/ICICT60155.2024.10544582

Riasat Khan

North South University

Proceedings of the 7th International Conference on Inventive Computation Technologies (ICICT 2024)
Xplore Part Number : CFP24F70-ART ; ISBN : 979-8-3503-5929-9

Prediction of Dengue Cases in Bangladesh using Explainable Machine Learning Approach

Sadia Sabrina Prome
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
sadia.prome@northsouth.edu

Tushar Basak
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
tushar.basak@northsouth.edu

Tasnim Islam Plabon
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
tasnim.plabon@northsouth.edu

Riasat Khan
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
riasat.khan@northsouth.edu

2024 International Conference on Inventive Computation Technologies (ICICT) | 979-8-3503-5929-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICICT60155.2024.10544582

***Abstract*—Dengue fever is transmitted through the wound of an Aedes mosquito that harbors the dengue virus. Transmission occurs when the mosquito acquires the virus from an individual with circulating dengue virus in their bloodstream during a biting interaction. Symptoms typically manifest within four to six days post-infection and can persist for about ten days. While some cases present mild symptoms that could be mistaken for the flu or other viral infections, the potential development of severe complications should not be overlooked. This study aims to develop a machine-learning model capable of utilizing relevant information to predict dengue outbreaks within the geographic regions of Bangladesh. For this purpose, a dataset from Bangladesh's weather forecast has been employed to predict dengue cases across 11 districts. Seven machine learning, including three ensemble learning techniques, have been applied in this work to forecast dengue cases and among these, the supper vector regression model with optimized hyperparameters achieved the best results, illustrating the lowest mean absolute error of 4.112. The prediction results obtained from the machine learning models have been analyzed using a Shapash-based explainable AI framework. Overall, the proposed data-driven analysis underscores the significant potential of machine learning algorithms in predicting dengue epidemics.**

***Index Terms*—Dengue, Explainable AI, Machine Learning, Prediction, Regression.**

I. Introduction

Dengue fever represents a major public health challenge in tropical and subtropical areas, as it is an infection transmitted by mosquitoes, including Bangladesh [1]. Due to its geographical location, favorable atmospheric conditions, dense population, and other conducive factors, our country is predisposed to a higher susceptibility to dengue [2]. Although dengue fever may remain asymptomatic or induce only mild manifestations, there exists the potential for more severe cases, occasionally culminating in fatality. Most individuals experiencing dengue exhibit mild or no symptoms, with recovery typically occurring within 1–2 weeks. Nevertheless, dengue can manifest as a severe illness in rare instances, posing a heightened mortality risk. In the event of symptom manifestation, they typically emerge within a window of 4–10 days post-infection, persisting for 2–7 days. These symptoms may encompass a high fever, intense headache, ocular pain, muscular and joint discomfort, nausea, vomiting, swollen glands, and a distinctive rash. The primary preventive measure is averting mosquito bites. Pain control often involves the use of acetaminophen paracetamol). At the same time, non-steroidal anti-inflammatory drugs like ibuprofen and aspirin are eschewed due to their potential to elevate the risk of bleeding. For those afflicted with severe dengue, hospitalization is frequently imperative.

Consequently, an urgent need is to formulate effective prevention, early detection, and control strategies for dengue. In 2023, dengue instilled a widespread sense of apprehension, with the current outbreak surpassing all previous records. In this work, the dataset has been obtained from the Bangladesh Meteorological and Health Services Departments. Our dataset incorporates essential variables such as rainfall, temperature, humidity, and the number of patients from each District in Bangladesh, which is crucial for predicting and understanding the dynamics of the ongoing dengue outbreak.

Significant works have been performed in dengue disease prediction and outbreak management. Dey et al. [3] predicted dengue cases for the ten months, i.e., August 2021 to May 2022, with a time series forecasting method. The investigation delved into examining the correlation between environmental factors such as temperature, rainfall, and humidity and the fluctuating trends of dengue cases in diverse cities in Bangladesh. The applied linear regression and SVR models attained average absolute error of 4.57 and 4.95, respectively. Gupta and colleagues [4] focused on developing a machine learning method to forecast dengue fever. The ensemble random forest technique accomplished the best accuracy of 87.21%. Dourjoy and colleagues [5] conducted a study utilizing machine learning to compare the effectiveness of SVM and Random Forest methods for forecasting dengue occurrences in Bangladesh.

According to the findings, SVM and RF produced 68% and 64% accuracies, respectively.

Mamum et al. [6] concentrated on forecasting the outbreak of dengue fever in Dhaka, Bangladesh, employing machine learning algorithms and considering various environmental factors. The authors achieved a high level of accuracy, attaining 97% accuracy with the SVM model for predicting dengue epidemics in Bangladesh. Rahman et al. [7] investigated the symptoms of dengue fever and forecast the likelihood of infection by applying machine learning techniques. The study leveraged four distinct machine learning algorithms to predict the occurrence of dengue fever based on observed symptoms. Habibur et al. [8] focused on predicting dengue fever using multiple machine learning algorithms, including boosted decision trees, Bayes point machine, a multiclass decision forest, etc. The study utilized a tenfold cross-validation method to evaluate their machine-learning model's effectiveness, attaining a significant accuracy rate of 95%.

Long short-term memory (LSTM), support vector regression, generalized boosted models, vector autoregression, and other techniques were used by Satya et al. [9] to forecast the dengue prevalence in Kerala, an Indian subcontinent state. The number of dengue cases was regarded as the target variable, and precipitation, mean temperature, relative humidity, soil moisture, and NINO3.4 were considered independent meteorological factors. The LSTM model showed better prediction performance with a higher R-squared value of 0.86 and a lower root mean square error of 0.345. Rustom et al. [10] introduced a machine learning model that combines Artificial Neural Networks ANN and SVM to provide the foundation of the conceptual framework. It had membership functions that explained if a diagnosis of dengue was positive or negative. A cloud storage system stores the patients' real-time data for later use. Remarkably, the suggested model demonstrated its efficacy in detecting dengue with an accuracy rate of 96.19%.

In this work, we aim to develop a predictive model that delivers highly accurate forecasts for dengue occurrences. To achieve this, we have delved into the intricate dynamics of dengue transmission, leveraging historical outbreak data and incorporating climatic variables of 10 years data (January 2011 to July 2021). Our approach involves utilizing various machine learning models, including powerful ensemble models, to uncover the most effective predictive outcomes. Furthermore, we have demonstrated Explainable AI methodologies to enhance our understanding of the model's decision-making processes, ensuring transparency and interpretability in our predictions.

II. Methodology

A. Workflow

In this work, the weather and dengue cases datasets are collected from Bangladesh Meteorological and Health Services, respectively. Necessary preprocessing techniques [11], feature scaling, missing value imputation, one hot encoding, etc., are applied. Employing the hold-out validation technique [12], we carefully split the dataset into training and testing sets of 90% and 10%, respectively. After optimizing hyperparameters, the best-optimized machine learning model is obtained. Finally, we evaluated the performance based on MSE, RMSE, MAE, and R2 scores [13]. Working sequences of the proposed dengue cases prediction system are illustrated in Fig. 1.

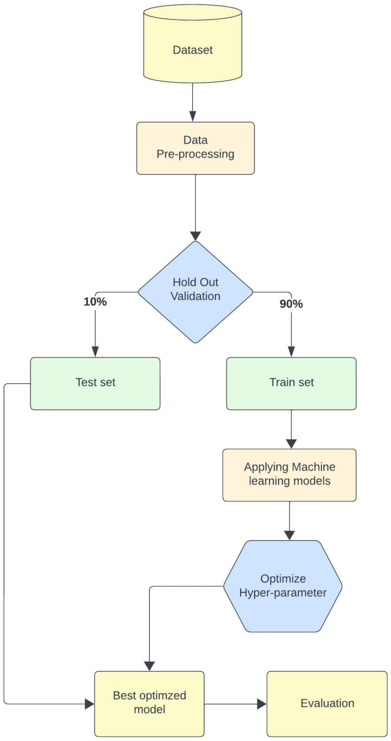

Fig. 1: Working procedure of the proposed dengue cases prediction system.

B. Dataset

A dataset comprising three climate factors, districts, months and dengue cases has been utilized in this work spanning from January 2011 to July 2021, providing valuable insights into dengue outbreaks in major cities in Bangladesh. The dataset comprises six key features: "Month," "District," "Maximum Temperature (C)," "Humidity (%)," "Rainfall (mm)," and "Number of dengue Patients Affected." With 1,397 data samples, the dataset captures dengue outbreaks' temporal and spatial dynamics over more than ten years.

C. Dataset Preprocessing

At first, we applied a label encoder [14] on two categorical features: month and district. Then, we applied Min-max scaling on numerical features, transforming elements by scaling each part to a given range, e.g., between zero and one. Next, we split the dataset into 90% for training and 10% for testing.

Fig. 2 and Fig. 3 show the individual humidity and maximum temperature bar charts, respectively. The figures illustrate that the most observed humidity and temperature are 33^0C and 90%, respectively.

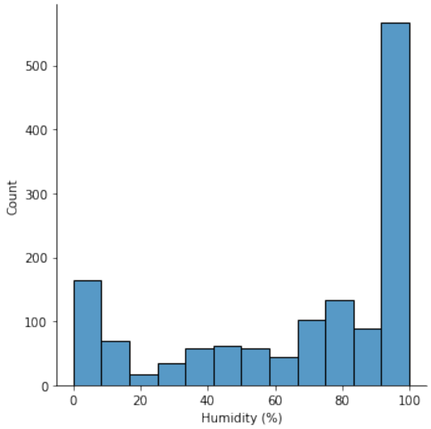

Fig. 2: Individual Barplot for Humidity (%).

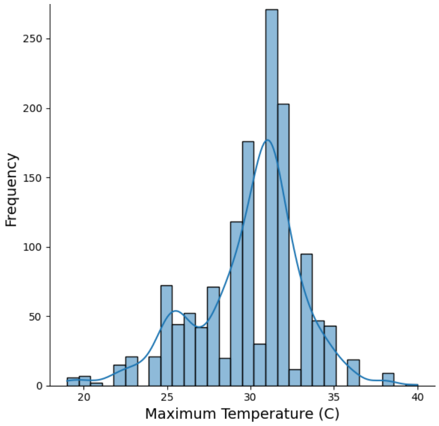

Fig. 3: Individual Barplot for Maximum Temperature (C).

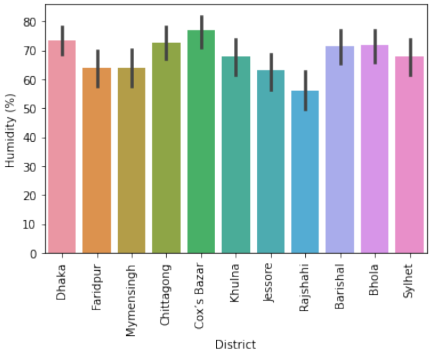

Fig. 4: Barplot for District vs. Humidity.

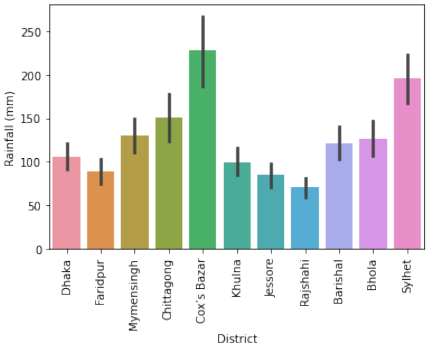

Fig. 5: Barplot for District vs. Rainfall.

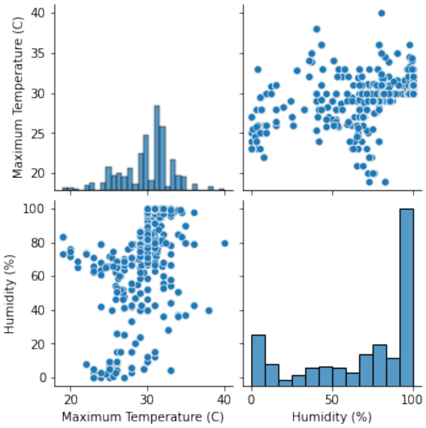

Fig. 6: Pair plot for Maximum Temperature vs. Humidity.

The relationship between district, humidity and rainfall is illustrated in Fig. 4 and Fig. 5. The correlation between maximum temperature and Humidity is depicted in Fig. 6.

The correlation between maximum temperature and Humidity is depicted in Fig. 6.

D. Machine Learning Models

We have used various machine learning models in this work, which has been briefly discussed in the following paragraphs.

Decision Tree: Decision Tree Regression is a predictive modeling method that uses a tree-like structure to make continuous numerical predictions based on the input data's features and interactions. It is useful for tasks where the target variable is a constant value, and it breaks down the data into subsets to estimate outcomes.

AdaBoost (Adaptive Boosting): It is an ensemble machine learning technique that combines multiple weak learners (typically decision trees) to create a more robust, more accurate model. It assigns different weights to misclassified samples during each iteration, emphasizing those that are harder to classify, and then combines the weak learners' outputs to make predictions.

XGBoost (Extreme Gradient Boosting): This optimized gradient boosting algorithm excels in predictive accuracy and speed. It uses an ensemble of decision trees, and its key in-

TABLE I: Performance metrics of the applied models with default hyperparamters

Model	**MSE**	**RMSE**	**R2**	**MAE**
Random Forest	733.779	27.088	2.0013	10.742
Decision Tree	1047.58	32.366	3.284	10.209
SVR	**259.625**	**16.112**	**0.0619**	**4.112**
Linear Regression	713.115	26.704	1.916	22.733
XGBoost	1178.290	34.326	3.819	11.915
AdaBoost	364.129	19.082	0.489	12.510
GPR	364.129	19.082	0.489	12.510

novations include regularization techniques, handling missing data, and parallel processing, making it a popular choice for various machine learning tasks, especially in structured data problems.

Random Forest: This ensemble machine learning technique builds several decision trees in the training phase and merges their outcomes to enhance precision and mitigate overfitting. This technique widely used for classification and regression tasks for its robustness, versatility, and ability to handle large datasets.

Gaussian Process Regression (GPR): It is a non-parametric, probabilistic approach to regression. It uses Gaussian processes, distributions over functions defined by mean and covariance functions. GPR models the relationship between input and output data, providing point estimates and uncertainty information. The choice of the kernel function is crucial in shaping the learned processes, and hyperparameters like lengthscale affect model behavior. GPR is useful for handling complex, non-linear, or uncertain relationships in data.

Support Vector Regression (SVR): SVR is a machine-learning technique for regression tasks. It identifies a hyperplane that best fits the data while minimizing errors within a defined margin. SVR is adequate for modeling complex relationships and handling outliers in data, making it valuable in predictive modeling with continuous numerical outcomes. The objective function of SVR is to minimize the error within a margin, subject to constraints.

III. Result Analysis

The experiment for the proposed dengue case forecasting system is conducted on a local machine, utilizing the specifications provided by Jupyter Notebook. Various Python libraries are employed, including Python version 3.9.13, NumPy (1.25.2), Pandas (1.4.4), and Matplotlib (3.6.2), to facilitate the experimentation process.

Performance metrics of the applied models with default hyperparameters for the proposed dengue case prediction system are illustrated in Table I. According to this table, the best-performing model is SVR, whose MAE is 4.112 and RMSE is 16.112. For optimized hyperparameters' the performance metrics of the applied models are shown in Table II. With its optimized hyperparameters, the SVR model achieved the best accuracy with MAE and RMSE of 3.865 and 16.04, respectively.

TABLE II: Performance metrics of the applied models with optimized hyperparameters

Model	**MSE**	**RMSE**	**R2**	**MAE**
Random Forest	536.924	23.171	1.196	7.842
Decision Tree	674.777	25.976	1.759	6.125
SVR	**257.290**	**16.04**	**0.052**	**3.865**
Linear Regression	713.115	26.704	1.916	22.733
XGBoost	574.179	23.962	1.348	10.205
AdaBoost	930.178	30.498	2.804	9.736
GPR	473.948	21.770	0.938	20.037

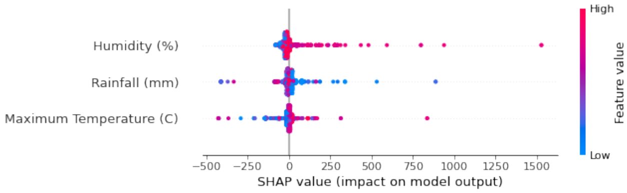

Fig. 7: Impact of various features in the dataset.

A. Explainable AI technique

An Explainable AI (XAI) technique that approximates the behavior of sophisticated machine learning models locally, giving interpretable insights into their predictions. It produces locally accurate and understandable explanations for each forecast. To visualize the impact of each feature of the employed dataset in the prediction, we have used the Shapash explainable AI framework. It is used to compute SHAP values for the training data, providing insights into the effects of each element on the model's predictions. Impacts of various features are shown in Fig. 7. As indicated by this figure, humidity and maximum temperature have a more significant impact on forecasting dengue cases than rainfall. This analysis is crucial for understanding the model's behavior and making informed decisions regarding the significance of each feature.

IV. Conclusion

Proactive monitoring of virus circulation in mosquitoes, coupled with the timely implementation of control measures, can mitigate the risk of widespread outbreaks, thereby preventing a significant toll on public health regarding sickness and mortality. Utilizing regression-based machine learning models and dataset comprising various environmental factors, our study aims to forecast dengue cases across 11 districts of Bangladesh. Through establishing correlations between climatic conditions and dengue occurrences, our findings yield valuable insights for implementing proactive measures. This research further supports authorities, healthcare organizations, and communities in comprehending and effectively responding to the patterns associated with the dengue epidemic. The applied SVR model with the corresponding optimized hyperparameters achieved the best performance concerning the lowest prediction errors. In the future, a comprehensive dataset with diverse features can be utilized.

REFERENCES

[1] R. Mahmood *et al.*, “Dengue outbreak 2019: Clinical and laboratory profiles of dengue virus infection in Dhaka city,” *Heliyon*, vol. 7, 2021.

[2] S. Hossain *et al.*, “Association of climate factors with dengue incidence in Bangladesh, Dhaka city: A count regression approach,” *Heliyon*, vol. 9, 2023.

[3] S. K. Dey, , *et al.*, “Prediction of dengue incidents using hospitalized patients, metrological and socio-economic data in Bangladesh: A machine learning approach,” *PLoS One*, vol. 17, 2022.

[4] G. Gupta *et al.*, “DDPM: A dengue disease prediction and diagnosis model using sentiment analysis and machine learning algorithms,” *Diagnostics*, vol. 13, p. 1093, 2023.

[5] S. M. K. Dourjoy, A. M. G. R. Rafi, Z. N. Tumpa, and M. Saifuzzaman, “A comparative study on prediction of dengue fever using machine learning algorithm,” in *Advances in Distributed Computing and Machine Learning*, pp. 501–510, 2021.

[6] M. T. Sarwar and M. Al Mamun, “Prediction of dengue using machine learning algorithms: Case study Dhaka,” in *International Conference on Electrical, Computer & Telecommunication Engineering*, pp. 1–6, 2022.

[7] T. Rahman and M. M. Rahman, “Evaluation of machine learning approaches for prediction of dengue fever,” in *Computer Networks and Inventive Communication Technologies*, pp. 165–175, 2023.

[8] M. Habibur Rahman, M. Omar Faroque, and F. S. Tithi, “Dengue fever prediction,” *Information and Communication Technology for Competitive Strategies*, p. 709–718, 2021.

[9] S. G. Kakarla *et al.*, “Weather integrated multiple machine learning models for prediction of dengue prevalence in India,” *International Journal of Biometeorology*, vol. 67, p. 285–297, 2022.

[10] 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,” *International Conference on Cyber Resilience*, 2022.

[11] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, “Diabetes prediction using machine learning and explainable AI techniques,” *Healthcare Technology Letters*, vol. 10, pp. 1–10, 2023.

[12] S. Solayman, S. A. Aumi, C. S. Mery, M. Mubassir, and R. Khan, “Automatic COVID-19 prediction using explainable machine learning techniques,” *International Journal of Cognitive Computing in Engineering*, vol. 4, pp. 36–46, 2023.

[13] N. E. J. Asha, E. U. Islam, and R. Khan, “Low-cost heart rate sensor and mental stress detection using machine learning,” in *International Conference on Trends in Electronics and Informatics*, pp. 1369–1374, 2021.

[14] M. N. I. Suvon, S. C. Siam, M. Ferdous, M. Alam, and R. Khan, “Masters and doctor of philosophy admission prediction of Bangladeshi students into different classes of universities,” *International Journal of Artificial Intelligence*, vol. 11, 2022.