Dengue Outbreak Prediction Using Simple Linear Regression

1st Nujat-E- Hasnat
Dept. of CSE
Independent University Bangladesh
Dhaka, Bangladesh
2330201@iub.edu.bd

2nd Sadman Sakib
Dept. of CSE
Independent University Bangladesh
Dhaka, Bangladesh
2221929@iub.edu.bd

3rd Md. Fazle Rabbi Mahmud
Dept. of CSE
Independent University Bangladesh
Dhaka, Bangladesh
2131412@iub.edu.bd

***Abstract*—Dengue fever is a viral mosquito borne illness that thrives in tropical areas, that has caused massive crises in the South Asian region of the world and among that Bangladesh has been facing several outbreaks throughout the years. The need for a prediction system is huge as caution is the best and possibly only line of defense against Dengue due to unavailability of a vaccine against the disease. Machine learning can help with this issue since it is very effective with its data driven capabilities and with linear regression is a much needed benchmark to aid the machine learning system. The main goal of this study is to predict the potential dengue outbreaks in Bangladesh based on weather variables by using the country's existing data and a simple linear regression model.**

The system primarily used four climate variables to predict future dengue cases and used the major performance metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to measure the efficiency of the system. The result of this implementation shows that of the four climate variables that were used in the system rainfall showed the best at predicting dengue cases since it had the best performance metric scores with a R^2 score of 0.9728 which suggests that the relation between dengue cases and rainfall is strong which the other variables lacked to prove as effective in predicting potential dengue outbreaks as well as the Rainfall data did.

***Index Terms*—Dengue, Prediction,Linear Regression, Outbreak, Bangladesh, Rainfall, MSE, RMSE, MAE.**

I. Introduction

Dengue fever, a viral mosquito-borne disease, is transmitted by Aedes mosquitoes prevalent in tropical and subtropical areas in the world. It has become a public health crisis worldwide, especially in South Asia over the past decades, and Bangladesh has emerged as the most affected country in recent years. In 2023, dengue became the most deadly epidemic in Bangladesh's history, and the capital's healthcare institutions were under pressure many times beyond their normal capacity.

According to the Directorate General of Health Services records (DGHS), the number of dengue cases nationwide exceeded 3 lakh, and over 1500 deaths due to the Aedes mosquito-borne tropical disease in the 2023 outbreak year, creating a groundbreaking severe epidemic history of Bangladesh [1]. In the Dhaka South City Corporation (DSCC) area, more than 50,000 people have been diagnosed with dengue, indicating the potential for people in densely populated areas to be more affected [2]. This recurring epidemic has not only constructed a public health concern in Bangladesh, but has also placed an enormous economic burden on the country's population management and workforce.

Machine learning models have proven effectiveness of disease prediction due to their data driven nature. Several models (LSTM, Random Forest, Gradient boosting) have been formulated and showed their accuracy. But linear regression remains an essential baseline due to their interpretability. This study focuses on predicted dengue cases based on weather data and evaluating their performance of linear regression.

II. Literature Review

Dengue is a major threat to the growing population of this world with its exponentially growing cases and reach throughout the world due to factors like climate change and urbanization. The necessity to predict future surges in dengue is very high due to the lack of vaccines or treatments to fight it, making readiness our main line of defense. Early detection and rapid access to medical care has shown a reduction in dengue driven mortality rate from 50% to around 2%. Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL) has become a very useful tool for creating advanced prediction systems by going beyond the pre existing linear and statistical methods [3].

The literature highlights a concerted shift toward complex non linear ML and DL models for predicting dengue outbreaks and incidence rates, in addition to models for individual diagnosis. For predictive capabilities, a comprehensive review of studies in Latin America, focusing on the One Health perspective classified dengue risk factors into four categories that are ecology of the vectors, serotypes, human conditions and environment, emphasizing that greater spatial and temporal resolution in predictive models is critically needed [4]. Studies focusing on outbreak or incidence rate prediction overwhelmingly leverage meteorological and medical data, including historical dengue incidence, temperature, rainfall, relative humidity and large-scale climatic guides like Nino3.4 [5]. Other researchers have successfully incorporated a broader range of data types, such as social media activity, mobile phone data on human mobility, satellite images for land use/ vegetation indices and vector indices to improve predictive power, recognizing the non linear, context specific and time variant nature of risk factors.

Regarding model performance in prediction ensemble ML methods and deep learning models have frequently demonstrated superior capability. A study in the Kerala state of India comparing statistical and ML/DL approaches found that the Long Short Term Memory (LSTM) deep learning model, particularly when using a lagged dataset incorporating optimal temporal lags, outperformed Vector Auto Regression (VAR), Generalized Boosted Regression (GBM) and Support Vector Regression (SVR), showing the lowest Root Mean Squared Error (RMSE) of 0.345 and highest $R\hat{2}$ of 0.86 [6]. Similarly, in predicting dengue transmission rates in Malaysia, ensemble methods like XGBoost, AdaBoost and Random Forest (RF) outperformed Logistic Regression, Naive Bayes, Decision Tree (DT) and Support Vector Machine (SVM) when leveraging both vector indices and meteorological data with XGBoost achieving the highest Area Under the Curve (AUC), accuracy and F1 score [7]. Another study in Semarang City, Indonesia found that the Extra Trees Classifier (ETC) achieved superior results for outbreak prediction with an accuracy of 89% and an AUROC of 95.2%.

Despite the strong performance of ensemble and deep learning models, simpler methods have shown contextual strength. Research in Selangor, Malaysia using climate variables to predict outbreaks, identified the SVM with a linear kernel as the best predictor, achieving an accuracy of 70% and a sensitivity of 63.54% after balancing the data and further highlighted that the "week-of-the-year" was the single most important predictor in the model [8]. LSTM is now given more importance due to its ability to model time series data and temporal dependencies, which is very important for a dengue prediction system.

A review of 32 articles found 48 specific ML and DL algorithms that were used to predict dengue based on patient symptoms and clinical data. The Support Vector Machine (SVM) was found to be the most effective algorithm in the studies that were reviewed, the second best was Random Forest (RF). The best performance model identified was a variation of the PCA-SVM algorithm which achieved an impressive 99.52% accuracy, 99.75% sensitivity and 99.09% specificity showing the true power of integrating dimensionality reduction techniques like the Principal Component Analysis (PCA).

Recent studies highlight the effectiveness of a hybrid model, but also mentions the superiority of simpler models for specific tasks. An analysis predicted dengue mortality rate in the Philippines, SVM Regression achieved a very high predictive accuracy and a correlation coefficient of 0.49 and the lowest Mean Absolute Error across all the models that were tested including RF and Linear Regression (LR) [10]. Comparatively for dengue cases in Malaysia a LSTM with Spatial Attentions (SSA-LSTM) model was the most efficient in achieving the lowest average Root Mean Square Error (RMSE) across various prediction lookback period, which outperformed benchmark models like SVM, DT and ANN.This proves the need for comparative studies since the optimal algorithms need a specific kind of dataset, prediction target and geographic information [11].

The recurring use of SVM and RF was seen in a lot of models both for prediction and diagnosis purposes, this was because SVM was very good at classifying non linear data while RF was good at reducing the overfitting issues which helped improve accuracy by combining multiple decision trees together. The number of different models that were successfully implemented highlights the fact that a single ML model is not best at this prediction compared to a hybrid system. A hybrid approach highlights the importance of testing multiple algorithms on specific integrated datasets which is necessary for creating a reliable prediction system.

III. Methodology

This system aims to find the relation between climate variables and dengue cases and henceforth finding the best climate variable that can be used to predict future dengue outbreaks. Single linear regression is used to create the model because of its transparent and straight forward approach to measuring influence for the effects of the climate features on dengue cases and providing easily understandable results for public stakeholders to view and make decisions on. Single linear regression is also more efficient in terms of computation and works well with limited datasets while also giving a strong benchmark for understanding patterns in the data.

A. Dataset Description

The dataset used in this project contains historical information related to dengue outbreaks [12]. It includes the number of dengue cases recorded over time along with several variables that may affect disease spread, such as temperature, rainfall, humidity and other environmental or seasonal indicators. These features help the model understand how changes in climate conditions relate to fluctuations in dengue cases.

Before training the model, the dataset was examined for missing or inconsistent entries. Some columns contained null values, which can cause errors during model training. To solve this, each data point in every year is added together so each year has one value with no null values instead of multiple data points which had the null values.

After cleaning the data, the dataset was divided into two parts. Eighty percent of the data was used for training the model, while the remaining twenty percent was reserved for testing. This split allowed the model to learn patterns from a large portion of the data while still leaving enough samples to evaluate its performance on unseen cases.

A linear regression model was selected for this project because it provides a clear way to analyze the relationship between the input variables and the number of dengue cases. It works well for predicting continuous values and helps show how each feature influences the final prediction. The cleaned and organized dataset served as the foundation for building this model, allowing it to generate predictions based on historical trends and environmental factors.

Overall, the dataset went through a complete preparation process that included handling missing values, splitting the data for training and testing, and selecting linear regression

as the prediction model. This process ensured reliable and consistent input for predicting future dengue case trends.

B. Data Preprocessing

The dataset used has monthly data for minimum temperature, maximum temperature, humidity, rainfall and dengue cases. To get clear values to work with, the data is added together to convert them from a monthly to a yearly standpoint. So let x be the set of monthly data points for each feature in a given year which is denoted by y, and let m be the number of months in the year for the next few equations.

For the Dengue Cases which are denoted by DC the annual number is calculated by adding all the dengue case values in a year.

$$DC(y) = \sum_{i \to x} DC_i$$

For Rainfall which is denoted RF the annual number is also calculated by adding all the values of rainfall in a year.

$$RF(y) = \sum_{i \to x} RF_i$$

For the Minimum Temperature which is denoted by MN the annual number is calculated by averaging all the values of minimum temperature in a given year.

$$MN(y) = \frac{1}{m} \sum_{i \to x} MN_i$$

For the Maximum Temperature which is denoted by MX the annual amount is also calculated by averaging the maximum temperature values for the year

$$MX(y) = \frac{1}{m} \sum_{i \to x} MX_i$$

And lastly for Humidity which is denoted by HM the same method of averaging all the data in the year is used to calculate the annual amount.

$$HM(y) = \frac{1}{m} \sum_{i \to x} HM_i$$

There is a difference in the way the four feature sets are aggregated is because unlike Dengue and Rainfall there cannot be cumulative values for temperatures and humidity so they minimum, maximum temperatures and humidity had to be averaged while the dengue and rainfall had to be summed up to get all the annual values.

C. Prediction Model with Linear Regression

A Linear Regressions model is implemented to find the relation between the four features and the dengue cases. For this let θ_1 is the intercept and θ_0 is the slope and the predicted dengue cases be denoted with y and X which is $x_1, x_2, x_3 and x_4$ where they are minimum temperature, maximum temperature, humidity and rainfall respectively for the equation.

$$y = \theta_0 + \theta_1 * X$$

D. Model Evaluation Metrics

The performance of the Linear Regression model was assessed using three standard regression metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). The performance of the system is measured with three standard metrics that are the Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE). For this dcA is the actual dengue cases and dcP is the predicted dengue cases and N is the number of data points in a year. For Mean Squared Error (MSE) it is calculated using the formula.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (dcA - dcP)^2$$

For Root Mean Squared Error (RMSE) it is calculated using the formula.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (dcA - dcP)^2}$$

For the Mean Absolute Error (MAE) it is calculated using the formula.

$$RMSE = \frac{1}{N} \sum_{i=1}^{N} |dcA - dcP|$$

IV. Result Analysis and Discussion

This project utilizes simple linear regression ($y = \theta_0 + \theta_1 X$) to evaluate the predictive capacity of four meteorological features: minimum temperature, maximum temperature, humidity and rainfall, on the annual incidence of Dengue cases. Performance was measured using the Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and the Coefficient of Determination R^2 [13]. The corresponding graphs are presented as figures for clarity and interpretation of the result.

The predictive performance metrics optimize over a defined range of θ_0 and θ_1.

A. Data Exploration and Trend Analysis

The initial analysis focused on understanding the behavior of dengue cases with respect to time and related environmental factors.As shown in Fig. 1, the visualization of dengue cases lets us know seasonal patterns and noticeable peaks due to those climatological data. Where we have summed the data of total dengue cases and rainfall. We have taken the mean value of temperature and humidity. These peaks in the graph suggests climatic conditions are favorable for mosquito breeding. It also indicates a strong relationship between dengue incidence and environmental variables. This observation justifies the inclusion of weather related data in the prediction model.

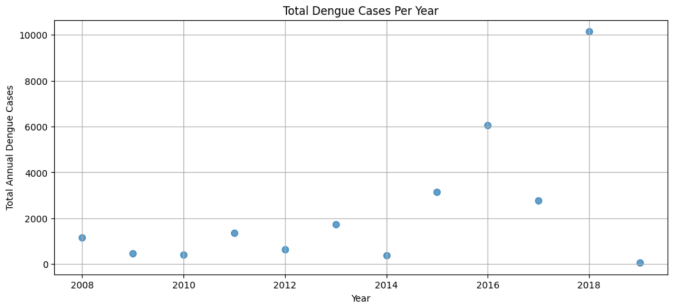

Fig. 1. Dengue Cases Over the year.

B. Feature Relationship Analysis

The plots shown in Fig. 2 demonstrate the relationship between dengue cases and input features such as temperature, rainfall and humidity. The common graphs indicates that higher dengue case counts are generally associated with increased rainfall. This confirms that these features play a significant role in influencing dengue outbreaks. This also mentions these data contribute meaningfully to the model's learning process.

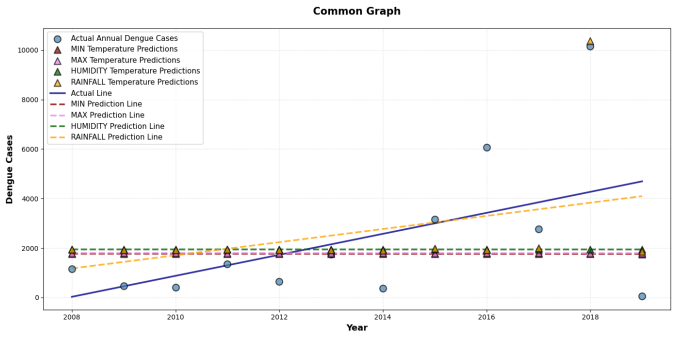

Fig. 2. Predicted lines over all features of model data.

C. Model Training Performance Evaluation

Model training performance is evaluated using the Root Mean Squared Error (RMSE). As shown in the Fig. 3 the RMSE value decreases during the training period. This indicates that our model effectively learns the underlying patterns from the data without significant overfitting. The stable and decreasing RMSE trend indicates stability and effectiveness of the chosen algorithm.

The trend set by the RMSE graph is also verifiable by the Mean Absolute Error (MAE) graph of the system as can be seen in Fig. 4. The graph here shows the same patter as of the RMSE further strengthening the viability of the chosen algorithm since even over the 50,000 iterations it used to calculating the both the RMSE and MAE the graph followed the same pattern and trend for both instances.

D. Prediction Accuracy and Evaluation

The prediction results are visualized in Table I and Table II, where actual dengue cases are compared against the predicted

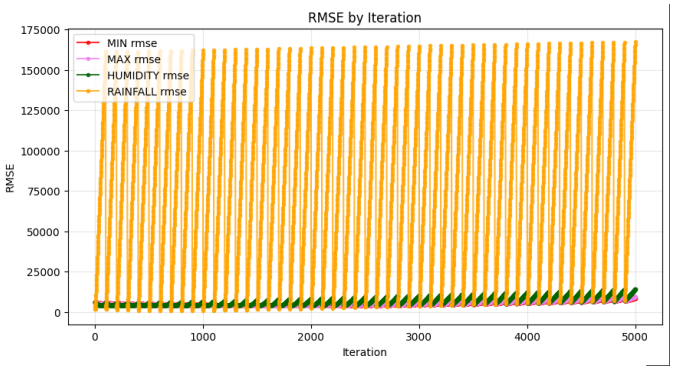

Fig. 3. RMSE by Iteration.

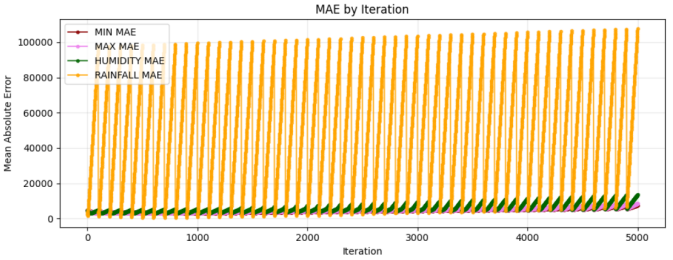

Fig. 4. MAE by Iteration.

values. In Table I we have observed train set metric and In Table II we have observed Test Set Metric. For the Temperature and Humidity the model showed poor performance with high error values of RMSE and MAE. The rainfall feature shows the highest accuracy of this model compared to the Temperature and Humidity feature.

TABLE I
TRAIN SET PERFORMANCE METRICS

Feature (X)	Parameter	Train Set Metric
Minimum Temperature	RMSE	3498.0524896528036
	MAE	3317.776636665913
	R^2	-2.6828244575443936
Maximum Temperature	RMSE	3588.539828778648
	MAE	3398.2338970654764
	R^2	-2.8758228276629363
Humidity	RMSE	3578.42046016853
	MAE	3389.0561225063093
	R^2	-2.8539946847395874
Rainfall	RMSE	1841.6613732440244
	MAE	1519.6681626181971
	R^2	-0.020817092995710107

The system used the data from the dataset and through linear regression predicted the dengue case for the coming year for all 4 climate variables in the dataset as can in the Fig. 5 (a), (b) and (c) all have similar outputs not really adhering to the the actual dataset plots hence why they are constantly off in their prediction all three models for minimum temperature, maximum temperature and humidity predicted that the upcoming years dengue case to be 3995, 4669 and 4665 for minimum

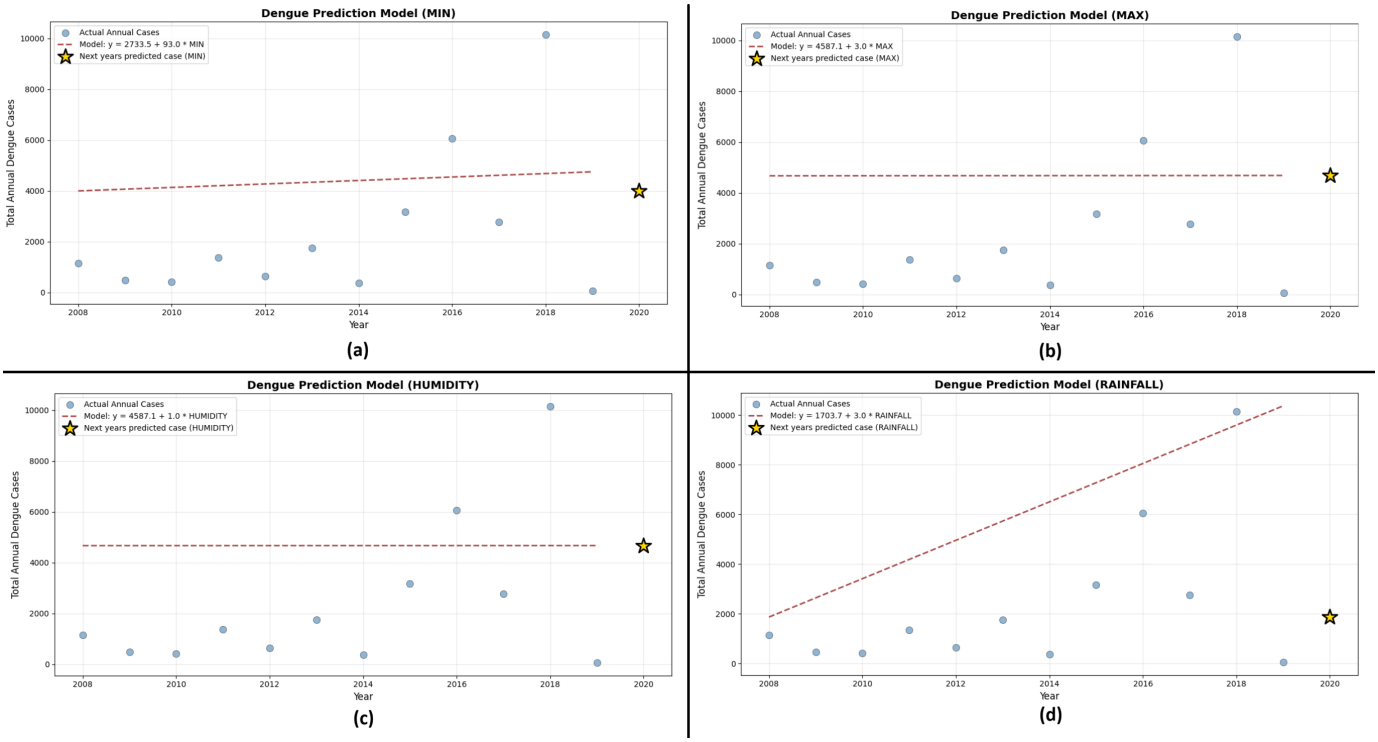

Fig. 5. a is the MIN graph, b is the MAX graph, c is the HUMIDITY graph and d is the RAINFALL graph.

TABLE II
TEST SET PERFORMANCE METRICS

Feature (X)	Parameter	Test Set Metric
Minimum Temperature	RMSE	3915.740460873288
	MAE	3642.759989090034
	R^2	0.00011856650939723323
Maximum Temperature	RMSE	3916.563738263403
	MAE	3636.287827772942
	R^2	-0.00030192423622232845
Humidity	RMSE	3916.3590643745865
	MAE	3632.216238481553
	R^2	-0.00019737833889355016
Rainfall	RMSE	645.5840235288154
	MAE	589.4983834712514
	R^2	0.9728214517296618

temperature, maximum temperature and humidity respectively but even the actual data in the dataset doesn't support those predictions since most years the actual dengue cases is usually way below the 4000 mark. Subsequently in (d) for the rainfall variable the model fit the actual data the best, so hence why it understood the pattern in the dataset the best which resulted in the best case prediction as can bee seen in the graph of a predicted dengue case of 1872 which is verifiable as that is in range of the actual yearly dengue case for most of the years in the dataset. So unless there are some anomalous variables in play the actual dengue cases should be around the predicted value provided by the model with the help of the rainfall data, further strengthening the bond between the quantity of dengue cases with rainfall.

E. Overall Discussion

Overall, the graphical results confirm that the proposed model performs effectively in predicting dengue outbreaks. The combination of environmental and historical data allows the model to detect meaningful patterns and seasonal trends. This analysis indicates that rainfall is the dominant factor and effective predictor for the annual dengue cases. Where as Temperature and Humidity are the very poor predictors for linear regression to predict annual dengue cases. The results highlight the model's potential as a decision support tool for health authorities to plan preventive measures and allocate resources more efficiently.

V. CONCLUSION

This study represents a machine learning approach that predict Dengue outbreaks in Bangladesh. Where the prediction is analyzed by the correlation between climatological variables such as Maximum Temperature, Minimum Temperature, Humidity, Rainfall and the incidence of Dengue cases per year. Using the Simple Linear Regression Model, we evaluated the individual impact of each weather factor on disease transmission. Our experimental results demonstrate that Rainfall is the most significant environment predictor for Dengue outbreaks in this region. The regression model based on Rainfall achieved an exceptionally high coefficient of determination (R^2), indicating a strong linear relationship between meteorological features and the number of Dengue cases. Although our model relying solely on Temperature

(Min/Max) or Humidity has yielded negligible predictive power when analyzed in isolation. This result suggests that while temperature and humidity likely play a role in mosquito breeding cycles, rainfall acts as the primary driver for creating the undynamic water sources necessary for rapid reproduction of a mosquito in Bangladesh. These findings provide critical insights for public health officials and also emphasize that rainfall monitoring should be central to early warning systems.

A. Future Works

While the current study establishes a strong baseline for correlation, several avenues exist to enhance the model's robustness and predictive accuracy in future iterations:

- **Implementation of Gradient Descent Optimization:** Currently, the model parameters (θ_0 and θ_1) were estimated using a manual iterative search. In future work, the Gradient Descent algorithm can be implemented to mathematically optimize these parameters. By adjusting the weights in each iteration to minimize the Cost Function, specifically the Mean Squared Error (MSE), the model can converge to the optimal solution faster and more accurately. This optimization is particularly important if the model is scaled to include more complex data.
- **Multivariate Linear Regression:** Since weather variables do not occur in isolation (e.g., high humidity often follows rainfall), future models can utilize a Multivariate Regression approach. This method allows Rainfall, Temperature, and Humidity to be combined in a single equation, enabling the analysis of their combined effects on Dengue incidence.
- **Non-Linear Modeling:** The relationship between climate and biological factors is rarely perfectly linear. Future work can explore non-linear models, such as Polynomial Regression or Support Vector Regression (SVR), to capture complex patterns. For example, thresholds of temperature that inhibit mosquito breeding can be identified.

These directions aim to enhance the performance and applicability of the proposed approach.

REFERENCES

[1] "2023 dengue outbreak in Bangladesh," Wikipedia, Sep. 28, 2023. https://en.wikipedia.org/wiki/2023_dengue_outbreak_in_Bangladesh

[2] "Dengue cases cross 50,000," The Business Standard, Oct. 06, 2025. https://www.tbsnews.net/bangladesh/dengue-cases-cross-50000-1254086 (accessed Dec. 13, 2025).

[3] K. Roster and F. A. Rodrigues, "Neural Networks for Dengue Prediction: A Systematic Review," arXiv (Cornell University), Jan. 2021, doi: https://doi.org/10.48550/arxiv.2106.12905.

[4] M. Cabrera, J. Leake, J. Naranjo-Torres, N. Valero, J. C. Cabrera, and A. J. Rodríguez-Morales, "Dengue Prediction in Latin America Using Machine Learning and the One Health Perspective: A Literature Review," Tropical Medicine and Infectious Disease, vol. 7, no. 10, p. 322, Oct. 2022, doi: https://doi.org/10.3390/tropicalmed7100322.

[5] D. Nur, Yu-Chuan (Jack) Li, C.-Y. Hsu, Muhammad Solihuddin Muhtar, and Hanif Pandu Suhito, "Artificial Intelligence Approach for Severe Dengue Early Warning System," Studies in health technology and informatics, Jan. 2024, doi: https://doi.org/10.3233/shti231091.

[6] S. G. Kakarla et al., "Weather integrated multiple machine learning models for prediction of dengue prevalence in India," International Journal of Biometeorology, pp. 1–13, Nov. 2022, doi: https://doi.org/10.1007/s00484-022-02405-z.

[7] Song Quan Ong, Pradeep Isawasan, A. Mohiddin, H. Shahar, Asmalia Md Lasim, and G. Nair, "Predicting dengue transmission rates by comparing different machine learning models with vector indices and meteorological data," Scientific Reports, vol. 13, no. 1, Nov. 2023, doi: https://doi.org/10.1038/s41598-023-46342-2.

[8] N. A. M. Salim et al., "Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques," Scientific Reports, vol. 11, no. 1, Jan. 2021, doi: https://doi.org/10.1038/s41598-020-79193-2.

[9] D. C. Andrade Girón, W. J. Marín Rodriguez, F. de M. Lioo-Jordan, and J. L. Ausejo Sánchez, "Machine Learning and Deep Learning Models for Dengue Diagnosis Prediction: A Systematic Review," Informatics, vol. 12, no. 1, p. 15, Feb. 2025, doi: https://doi.org/10.3390/informatics12010015.

[10] A. Saeed, N. Muhammad, Muhammad Sauood, S. N. Ali, and N. Hussain, "PREDICTION OF DENGUE CASES AND DEATHS USING MACHINE LEARNING ALGORITHM," Pakistan Journal of Scientific Research, vol. 3, no. 2, pp. 210–216, May 2024, doi: https://doi.org/10.57041/pjosr.v3i2.1042.

[11] M. A. Majeed, Z. Mohd, Zed Zulkafli, and Aimrun Wayayok, "A Deep Learning Approach for Dengue Fever Prediction in Malaysia Using LSTM with Spatial Attention," International Journal of Environmental Research and Public Health, vol. 20, no. 5, pp. 4130–4130, Feb. 2023, doi: https://doi.org/10.3390/ijerph20054130.

[12] "Dengue Incidents and Weather Data of Bangladesh," www.kaggle.com. https://www.kaggle.com/datasets/fazlyrabbi/dengue-incidents-weather-of-bangladesh

[13] "Google Colab Notebook for Dengue Outbreak Prediction Model," Google Colaboratory, 2025. https://colab.research.google.com/drive/1F2kpTTDDUHsRpHWoOiIoLZek9KZKkmh8

[14] Khokon Kanti Bhowmik, J. Ferdous, Prodip Kumar Baral, and Mohammad Safiqul Islam, "Recent outbreak of dengue in Bangladesh: A threat to public health," Health science reports, vol. 6, no. 4, Apr. 2023, doi: https://doi.org/10.1002/hsr2.1210.