Environmental Drivers of Local Dengue Outbreaks: A Machine Learning Approach

Yash Parab
School of Computer Science
University of Nottingham
Nottingham, UK, NG8 1BB
yashparab05@gmail.com

Abdullah Ansari
School of Computer Science
University of Nottingham
Nottingham, UK, NG8 1BB
abd2452001@gmail.com

Sarvesh Shantanu
School of Computer Science
University of Nottingham
Nottingham, UK, NG8 1BB
sarveshshantanu1@gmail.com

***Abstract*— This research investigates advanced machine learning approaches for accurately predicting local epidemics of dengue fever, a significant public health challenge intensified by climate change. By leveraging historical dengue surveillance data and environmental variables from Iquitos, Peru, and San Juan, Puerto Rico, the study aims to develop robust predictive models capable of forecasting weekly dengue case incidence. The methodology encompasses data preprocessing techniques, including handling outliers and scaling features, followed by model fitting using algorithms like K-Nearest Neighbors, Decision Trees, Random Forest, and XGBoost. The results demonstrate that ensemble methods, particularly XGBoost, outperformed individual models, with further improvements observed after preprocessing the data. Feature importance analysis highlights the significance of environmental factors like temperature, precipitation, and vegetation indices in dengue transmission dynamics. The findings have practical implications for public health authorities, enabling early warning systems and targeted interventions to mitigate the impact of dengue outbreaks. Future research directions include exploring the generalizability of other regions, incorporating socioeconomic factors, and investigating advanced deep-learning techniques for dengue forecasting.**

Keywords: Dengue fever, epidemic prediction, ensemble techniques, Environmental factors, public health, climate change, advance machine learning techniques, socioeconomic factors, predictive modelling.

I. Introduction

The potential for severe sickness and widespread transmission of dengue fever, a mosquito-borne disease common in tropical and sub-tropical locations worldwide, presents a substantial challenge to public health systems. Dengue transmission dynamics are closely associated with climate variables, including temperature and precipitation, and the global spread of the disease is intensified by climate change. As the incidence of dengue fever escalates, particularly in Latin America where nearly half a billion cases are reported annually, there is a pressing need for innovative approaches to predict and mitigate the impact of dengue epidemics. Recognizing this imperative, several U.S. Federal Government departments, including the Department of Health and Human Services, Department of Defence, and Department of Commerce, have collaborated under the auspices of the National Science and Technology Council (NSTC) to design an infectious disease forecasting project specifically targeting local epidemics of dengue. Aligned with former US President Barak Obama's vision of fostering collaboration and data-intensive discoveries, Predict the Next Pandemic (PtNP) Initiative, initiated by Dr. John Holdren, aims to harness predictive capabilities to anticipate disease emergence and forecast the progression of infectious diseases. Through the establishment of the Pandemic Prediction and Forecasting Science and Technology (PPFST) Working Group, the initiative seeks to mobilize the academic community and the public towards leveraging big data for infectious disease forecasting. Dengue fever affects residents of the tropical regions of the U.S. like Puerto Rico. Experts estimate that around 390 million dengue infections occur worldwide each year, including about 500,000 severe cases mostly children requiring hospitalization. Case counts are climbing as the disease moves into new areas. [1] [2]

Our research attempts to address the pressing challenge of predicting local epidemics of dengue fever in regions such as Iquitos, Peru, and San Juan, Puerto Rico. By leveraging historical dengue surveillance data and a comprehensive set of environmental variables, including temperature, health and density of vegetation, precipitation, and humidity. To ensure the robustness and accuracy of our forecasting models, we aim to address potential outliers and apply standardization techniques to the dataset. By addressing these issues, we anticipate improved predictive accuracy and model performance.

Our research questions are as follows:

1. Can addressing outliers in the data and applying scaling techniques improve the performance and predictive accuracy of the machine learning models?

2. How do ensemble methods compare to other algorithms in predicting dengue cases accurately?

These research questions are appropriate and relevant to the context of the dataset.

We aim to develop advanced machine learning models capable of accurately forecasting the weekly incidence of dengue cases. Specifically, we will employ Random Forest Machine Learning algorithm to capture the complex relationships between environmental conditions and dengue transmission dynamics. Our novel approach has the potential to provide early warning systems for local health authorities, enabling timely interventions and resource allocation to mitigate the impact of dengue outbreaks.

II. Literature Review

"Predicting Dengue Incidence in Iquitos, Peru" by Johansson et al. (2009): This study focused on predicting dengue incidence in Iquitos, Peru, using climate data and

historical dengue surveillance data. The authors employed statistical models to assess the relationship between climate variables and dengue transmission dynamics. Their findings underscored the importance of temperature and precipitation in dengue incidence prediction. [3]

"Dengue confirmed-cases prediction: A neural network model" by Aburas et al. (2010) aimed to forecast (Back to the Future: Using Historical Dengue Data to Predict the Next Epidemic., 2015)confirmed cases using Artificial Neural Networks based on six years of recorded data. Their study identified mean temperature, mean relative humidity, total rainfall, and the total number of dengue confirmed-cases as crucial features for predicting the number of dengue confirmed-cases. [4]

"DDPM: A Dengue Disease Prediction and Diagnosis Model Using Sentiment Analysis and Machine Learning Algorithms" by Gupta et al. (2023) emphasized optimizing machine learning techniques for disease diagnosis, particularly in the context of dengue fever. They highlighted the potential of reason-based models and data decentralization to enhance healthcare outcomes while reducing costs. [5]

"Dengue Prediction using Machine Learning Algorithms" by Sarma et al. (2020) proposed an ML model to predict dengue fever using patient data collected from medical college hospitals in Dhaka and Chittagong. They pre-processed the data into 23 features and applied Decision Tree (DT) and Random Forest (RF) algorithms to classify three types of dengue fever. [6]

" Smart System for Dengue Fever Diagnosis: A Machine Learning Approach" by AL-Hagree et al. (2023) utilized machine-learning techniques commonly employed in the medical field for disease diagnosis. Their study highlighted the effectiveness of decision tree algorithms in accurately classifying different types of dengue fever. They proposed developing an Android application leveraging decision tree algorithms to facilitate rapid and accurate diagnosis of dengue fever. [7]

Tabel I: Brief Overview of related Research Papers regarding Dengue Disease Outbreak

Subject of Research	Algorithm	Prediction Accuracy	Author 's name
Dengue confirmed-cases prediction: A neural network model	Artificial Neural Networks (ANN)	82.39%	Aburas, H.M.; Cetiner, B.G.
DDPM: A Dengue Disease Prediction and Diagnosis using ML Algorithms	Random Forest (RF)	87.2%	Gupta, G., Khan, S.
Dengue Prediction using ML Algorithms	Decision Tree (DT)	79%	Sarma, S. Hossain , T.
Smart System for Dengue Fever Diagnosis: A ML Approach	Decision Tree (DT)	93.7%	S. AL-Hagree

Table I provides a concise overview of research papers related to Dengue Disease Outbreak, summarizing key information such as the subject of research, algorithms used, prediction accuracy, and authors. These findings highlight the diversity of approaches and techniques employed in predicting and diagnosing Dengue Fever, showcasing the importance of machine learning and data analysis in addressing public health challenges.

III. Methodology

The methodology in our research encompasses several key stages which includes Data Sources, Data Collection, Exploratory Data Analysis (EDA), Data Preprocessing, Feature Selection, Model Fitting and Evaluation as shown in Figure 1.

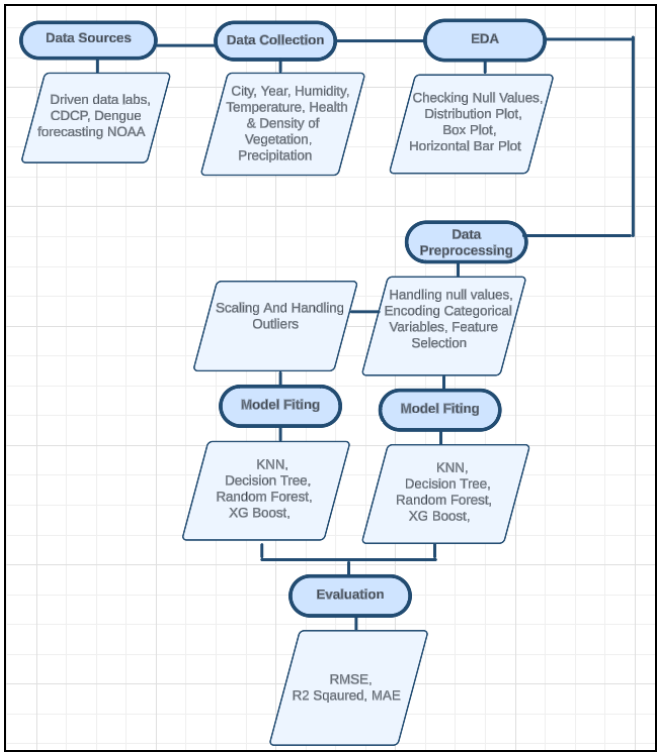


Figure I: Flow chart of Dengue Pediction Model

A. Data Collection

To achieve forecasting of weekly incidence of dengue cases, we employ a comprehensive data-driven approach that leverages historical dengue surveillance data from year 1990 to 2010 and a wide range of environmental variables. This study uses data from the DengAI competition (open data of dengue illness competition: DengAI: Predicting Disease Spread (drivendata.org). The data was provided by the National Science and Technology Council (NSTC) as part of the Predict the Next Pandemic (PtNP) Initiative.

The dataset used in this research comprises three main files:

1. Dengue Features Training file; consists of various environmental features such as temperature, health and density of vegetation, precipitation, and humidity.

2. Dengue Labels Training file; consists of city, year, week of year and total cases in a particular week.

3. Dengue Feature Test file; contains similar data features present in Dengue Features Training file.

Table II: Data Distribution

Labelled Data	Dataset Type	Data
Dengue Features	Training	1456
Dengue Labels	Training	1456
Dengue Features	Test	416

B. *Exploratory Data Analysis*

Exploratory data analysis (EDA) is a crucial step in any data-driven research project, as it provides insights into the characteristics and quality of the dataset. In this study, we performed EDA to identify potential issues, such as missing values and outliers, and to gain a better understanding of the data distribution.

Checking Missing Values: One of the first steps in EDA is to check for missing values in the dataset. The output revealed several features like temperature, health and density of vegetation, precipitation, and humidity consisting of null values as shown in Figure 2.

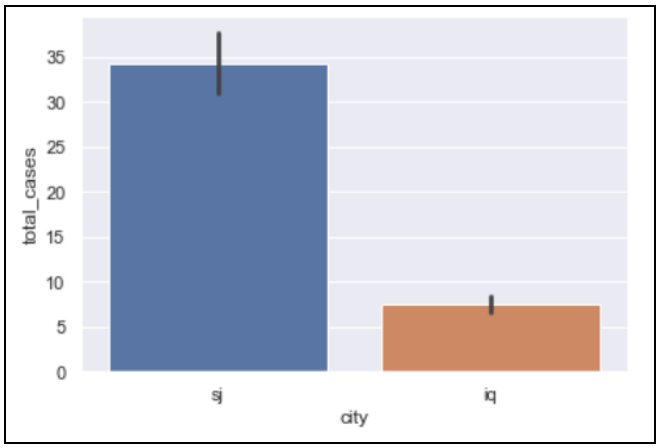
FIGURE 2: NULL VALUES COUNT

ndvi_ne	194
ndvi_nw	52
ndvi_se	22
ndvi_sw	22
precipitation_amt_mm	13
reanalysis_air_temp_k	10
reanalysis_avg_temp_k	10
reanalysis_dew_point_temp_k	10
reanalysis_max_air_temp_k	10
reanalysis_min_air_temp_k	10
reanalysis_precip_amt_kg_per_m2	10
reanalysis_relative_humidity_percent	10
reanalysis_sat_precip_amt_mm	13
reanalysis_specific_humidity_g_per_kg	10
reanalysis_tdtr_k	10
station_avg_temp_c	43
station_diur_temp_rng_c	43
station_max_temp_c	20
station_min_temp_c	14
station_precip_mm	22

Outlier Detection: Initially, plotted the distribution plot on all the features in order to check whether the features follow a Normal Distribution. We then diagnosed Box Plots on all the features to detect the outliers.

We also explored and visualized the Total number of cases for both the cities in our dataset.

FIGURE 3: TOTAL NUMBER OF DENGUE CASES PER CITY



C. *Data Preprocessing*

Data preprocessing is a crucial step in machine learning pipelines to ensure the quality and suitability of the data for modeling. In this study, we performed several preprocessing techniques to handle missing values, encode categorical variables, and standardize numerical features.

Handling Missing/Null Values: To address the missing values identified during EDA, we employed the Simple Imputer class from scikit-learn. We then used the mean imputation strategy to fill in the missing values for numerical features.

Encoding Categorical Variables: The dataset contained categorical variable as 'city' which cannot be directly used in machine learning algorithms. We utilized the Map class from Python to convert the categorical variables into numerical. We mapped the city San Juan as 0 and Iquitos as 1.

Handling Outliers: Outliers are data points that significantly deviate from the rest of the observations in a dataset. They can have a substantial impact on the performance of machine learning models, leading to biased results or poor generalization.

In our study, we visualized box plots for each feature to identify the presence of outliers. After detecting the presence of outliers in several features, we employed a robust technique known as the Interquartile Range (IQR) method to cap(limit) the outlier values.

The IQR method is based on the following calculations:

i. Calculate the first quartile Q1 (25th percentile) and third quartile Q3 (75th percentile) of the feature distribution.

ii. Compute the Interquartile Range (IQR) as IQR = Q3 - Q1.

iii. Define the lower and upper limits for outlier detection:

Lower limit = Q1 - 1.5 * IQR

Upper limit = Q3 + 1.5 * IQR

By capping the values lower than 25th percentile and values higher than 75th percentile to 25th and 75th percentile respectively we limited outliers influence on the machine learning models, potentially improving their performance and generalization capabilities.

Standardizing Numerical Features: Many machine learning algorithms can benefit from standardizing numerical features, as it ensures that all features are on a similar scale and prevents features with larger values from dominating the model's predictions. We employed the StandardScaler class from scikit-learn to standardize the numerical features in the dataset.

D. Feature Selection

Feature selection reduces the number of input variables when developing a predictive machine leaning model. It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and in some cases to improve the performance of the model [6]. In our data set we dropped the column week_start_date. As the name suggests, the column displays the starting date of the week, which does not give us any significance in our further analysis. Also, with regards to the same we have similar features such as 'year' and 'weekofyear'.

Analyzing Important Features: To gain insights into the relative importance of different features in predicting dengue cases, we employed the ExtraTreesClassifier algorithm from scikit-learn. This algorithm is an ensemble of decision trees that can provide an estimate of feature importance based on the decrease in impurity, Gini impurity or entropy when splitting on a particular feature.

Gini impurity and entropy are both measures used to evaluate the impurity or uncertainty of a node in a decision tree. These measures are used to determine the quality of a split during the tree construction process. The goal is to find the split that maximizes the purity of the resulting child nodes.

Gini Impurity: It is a measure of the probability of misclassifying a randomly chosen element in a dataset if it were randomly labelled according to the distribution of class labels in the subset. The Gini impurity for a node is calculated as follows:

$$Gini(node) = 1 - \Sigma(p\_i) \hat{} 2$$

where,

p_i is the proportion of instances belonging to class i in the node.

During the construction of each decision tree, the algorithm calculates the Gini impurity for each feature and selects the feature that provides the maximum decrease in impurity for the split. This process is repeated recursively until the tree is fully grown or a stopping criterion is met.

E. Model Fitting

To address our research questions and investigate the impact of outlier handling and scaling techniques on the performance and predictive accuracy of machine learning models, we employed a comprehensive modeling approach. Initially, the dataset was split into training and testing sets, with an 80:20 ratio, to ensure proper evaluation of the models' generalization capabilities.

1. Baseline Models: As a baseline, we trained several machine learning algorithms on the cleaned dataset without addressing outliers or applying scaling techniques.

The following models were considered:

i. K-Nearest Neighbours (KNN): A non-parametric algorithm that predicts the target variable based on the similarity of the input features to its neighbours in the training set.

For regression, 'i' in K nearest neighbors the prediction is given by:

$$y\_pred = (1/K) * \Sigma(y\_i)$$

where,

y_i is the target value of the i-th nearest neighbour.

ii. Decision Tree: A tree-based model that learns decision rules from the data features to predict the target variable.

For a node with classes C_1, C_2, ..., C_k, the Gini impurity is given by:

$$Gini(node) = 1 - \Sigma(p\_i) \hat{} 2$$

where,

p_i is the proportion of instances belonging to class C_i in the node.

iii. Random Forest: An ensemble learning method that constructs multiple decision trees and combines their predictions to improve accuracy and robustness, particularly for the feature importance calculation, can be expressed as follows:

$$RF_{fi}(i) = \Sigma^{T}_{j=1} normfi_{sub}(ij)/T$$

where,

$RF_{fi}(i)$: is the importance of feature (i) calculated from all trees in the Random Forest model.

$normfi_{sub}(ij)$ is the normalized feature importance for feature (i) in tree (j).

T is the total number of trees in the forest

iv. XGBoost: A gradient boosting algorithm that builds an ensemble of weak prediction models, typically decision trees, in a sequential manner to achieve high predictive performance.

The objective function optimized by XGBoost is:

$$Obj = \Sigma (L (y\_i, y\_i\_pred)) + \Sigma(\Omega(f\_i))$$

where,

L is the loss function (e.g., logistic loss for classification, mean squared error for regression),

y_i is the true target value,

y_i_pred is the predicted value,

Ω(f_i) is the regularization term that controls the complexity of the model.

2. Outlier Handling and Scaling: To investigate the potential impact of outlier handling and scaling techniques, we preprocessed the dataset and after preprocessing the dataset, we trained the same set of machine learning

algorithms (KNN, Decision Tree, Random Forest, and XGBoost) on the preprocessed data. The performance of all trained models was evaluated using appropriate metrics, such as:

1. Root Mean Squared Error (RMSE): It measures the square root of the mean squared differences between predicted and actual values. Lower RMSE indicates better model accuracy.
2. R-squared (R^2): Represents how much of the variance in the target variable is explained by the model's inputs. Values range from 0 to 1, with higher values indicating a better model fit.
3. Mean Absolute Error (MAE): Calculates the average magnitude of errors in the predictions, ignoring their direction. Lower MAE signifies better model performance.

IV. RESULTS

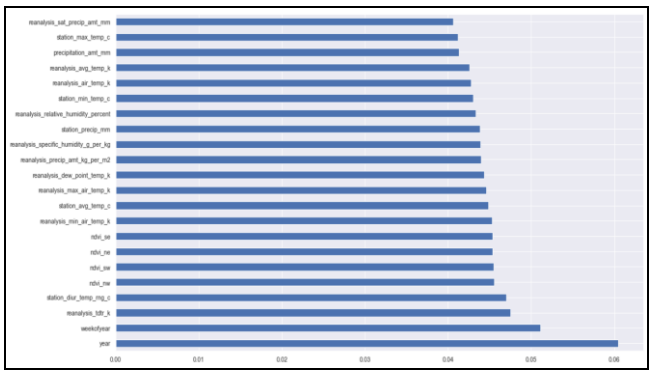
This section presents the outcomes obtained at various stages of the methodological approach employed in this study. It provides a comprehensive overview of the findings resulting from the different techniques and analyses conducted.

1. EDA: The exploratory data analysis (EDA) stage played a crucial role in understanding the characteristics and quality of the dataset. The distribution plots revealed that most of the features did not follow a normal distribution, indicating the presence of potential outliers or skewness in the data. The box plots provided visual confirmation of the existence of outliers across multiple features. Our Bar plot analysis revealed that city San Juan has nearly twice as much dengue cases that Iquitos.

2. Data Preprocessing: By eliminating missing/null values in all the features, encoding the categorical feature 'city', capping the outliers present in all the features, scaling the numerical features to ensure each feature contribute equally to the analysis we obtained a cleaned dataset that can be further used for fitting advance machine learning models.

3. Feature Selection: The feature selection process provided valuable insights into the relative importance of different features in predicting dengue cases.

FIGURE 4: ANALYSIS OF IMPORTANT FEATURES



By analyzing the results obtained in Figure 4, we could identify 'year', 'weekofyear' and 'Diurnal temperature range' as the most important features in our dataset and we could also depict that all other features share equal importance for model fitting.

4. Model Fitting: In this study, we explored various machine learning algorithms to develop accurate predictive models for forecasting the weekly incidence of dengue cases. The models were trained and evaluated using the preprocessed dataset.

The performance of the trained models was assessed using three widely adopted evaluation metrics such as mean absolute error (MAE), root mean squared error (RMSE), and R-squared (R^2) for regression tasks, or accuracy.

In the analysis mentioned in below table 3 one of our team member performed model fitting without scaling and handling outliers, with regards to that we assessed the performance of the trained models.

Table III: Data Distribution

Without Handling Outliers and Scaling			
Models	RMSE	R^2	MAE
KNN	38.4886	0.0853	19.8253
Decision Tree	24.6556	0.6246	12.9965
Random Forest	24.9076	0.6169	11.7750
XG Boost	19.9990	0.7530	10.1922

Similar such analysis was performed by remaining two members of the group to study the model fitting by handling the outliers and scaling the features and further assessed the performance of the trained models as shown in Table 4.

Table IV: Data Distribution

With Handling Outliers and Scaling			
Models	RMSE	R^2	MAE
KNN	38.6367	0.0782	19.6979
Decision Tree	27.7233	0.5254	13.4383
Random Forest	24.5887	0.6266	11.7102
XG Boost	17.5314	0.8102	9.3457

Amongst the various models used for fitting the dataset we observed that XG Boost outperformed other models.

V. DISCUSSION

The main objective of our study was to create sophisticated machine learning models that can reliably predict the weekly occurrence of dengue cases in Iquitos,

Peru, and San Juan, Puerto Rico. Utilizing past dengue surveillance data alongside various environmental factors, we aimed to develop novel strategies to forecast and manage the effects of dengue outbreaks in these areas.

With respect to our research questions mentioned above, we answer those below:

1. Addressing outliers and scaling features generally improved model performance, especially for ensemble methods like Random Forest and XGBoost. After outlier handling and scaling, XGBoost showed the most significant improvements with lower RMSE, MAE, and higher R^2.
2. Ensemble methods outperformed individual models like KNN and Decision Trees in predicting dengue cases accurately. Without outlier handling and scaling, XGBoost performed best, followed by Random Forest. XGBoost remained the top performer with outlier handling and scaling, but Random Forest outperformed KNN and Decision Trees.

In summary, ensemble methods like Random Forest and XGBoost were more accurate than individual models, with XGBoost being the most robust and accurate overall, particularly after addressing outliers and scaling features.

Furthermore, our findings align with previous studies that have highlighted the significance of temperature, precipitation, and vegetation indices in predicting dengue incidence. The feature importance analysis conducted in our study reinforced the relevance of these environmental factors, providing valuable insights for public health authorities and policymakers in devising targeted interventions and resource allocation strategies.

While our research has yielded promising results, it is essential to acknowledge its limitations. First, our study focused on specific regions, and the generalizability of our findings to other geographical areas may be limited due to variations in local climatic conditions and dengue transmission patterns. Additionally, the availability and quality of historical data can impact the performance of predictive models, underscoring the need for robust surveillance systems and data collection practices.

In conclusion, our study demonstrates the potential of advanced machine learning techniques, coupled with rigorous data preprocessing, in accurately forecasting the weekly incidence of dengue cases.

VI. CONCLUSION AND FUTIRE WORK

Our research has showcased the remarkable capabilities of sophisticated machine learning approaches, especially ensemble techniques like Random Forest and XGBoost, in precisely predicting the weekly dengue case counts in Iquitos, Peru, and San Juan, Puerto Rico. By harnessing historical dengue surveillance data coupled with a comprehensive array of environmental factors, this study has shed light on the intricate relationships between climatic variables and dengue transmission dynamics. The findings unveil valuable insights that can guide targeted interventions and inform strategic planning to combat the public health challenges posed by dengue outbreaks in these regions.

The practical utility of our findings lies in their ability to inform public health authorities and policymakers in devising targeted interventions and resource allocation strategies to mitigate the impact of dengue epidemics in these regions. By providing early warning systems and accurate forecasts, our models can facilitate timely and effective measures to prevent and control the spread of dengue fever, ultimately contributing to improved public health outcomes.

While our research has yielded promising results, there remain opportunities for further exploration and advancement in this field. Future studies could investigate the generalizability of our findings to other geographical areas, as local climatic conditions and dengue transmission patterns may vary. Additionally, incorporating socioeconomic factors, population dynamics, and vector control measures into the predictive models could provide a more comprehensive understanding of dengue transmission dynamics.

As machine learning techniques continue to evolve, exploring the potential of deep learning models and advanced neural network architectures in dengue forecasting could yield valuable insights. To optimize the performance of each machine learning model, we could employ hyperparameter tuning techniques. Specifically, using combinations of manual tuning and grid search cross-validation to explore different hyperparameter configurations and identify the optimal settings for each algorithm.

VII. REFERENCES

[1] "Back to the Future: Using Historical Dengue Data to Predict the Next Epidemic.," *The White House,* p. 1, 2015.

[2] L. P. Campbell, C. Luther, D. Moo-Llanes, J. M. Ramsey, R. Danis-Lozano and A. T. Peterson, "Climate change influences on global distributions of dengue and chikungunya virus vectors.," *Philosophical Transactions of the Royal Society B: Biological Sciences, 370(1665), 20140135,* p. 9, 2015.

[3] T. JM, C. R, L. P, B. LC and P. T, "A sailor's pain: Veterans' musculoskeletal disorders, chronic pain, and disability.," *Can Fam Physician,* p. 4, 2009.

[4] H. Aburas, B. Cetiner and M. Sari, "Dengue confirmed-cases prediction: A neural network model," *Expert Syst. Appl.,* p. 7, 2010.

[5] G. K. S. G. V. A. A. A. B. I. S. T. A. B. M. A. S. A. &. A.-s. M. Gupta, "DDPM: A Dengue Disease Prediction and Diagnosis Model Using Sentiment Analysis and Machine Learning Algorithms.," 2023.

[6] D. Sarma, S. Hossain, T. Mittra, M. A. M. Bhuiya, I. Saha and R. Chakma, "Dengue Prediction using Machine Learning Algorithms," IEEE 8th R10 Humanitarian Technology Conference (R10-HTC), Kuching, Malaysia, 2020.

[7] S. AL-Hagree, "Smart System for Dengue Fever Diagnosis: A Machine Learning Approach," 3rd International Conference on Emerging Smart Technologies and Applications (eSmarTA), Taiz, Yemen, 2023.