Climatologically-driven temporal predictive modeling of dengue cases in Philippine locations

Dela Cruz, Joshua
College of Engineering
University of the Philippines Diliman
Quezon City, Philippines
jrdelacruz12@up.edu.ph

Dela Cruz, Mary Nathalie
College of Engineering
University of the Philippines Diliman
Quezon City, Philippines
mgdelacruz12@up.edu.ph

***Abstract*—Dengue fever is a major public health concern, particularly in countries like the Philippines where dengue incidence is surging and highly variable. To address this, predictive models for dengue cases were developed for Bulacan, Quezon City, and Rizal—Philippine locations within the same dengue case cluster—using statistical and machine learning techniques. The process started with K-means clustering and t-SNE for location grouping, followed by stationarity checks to ensure data stability. The study integrated temporal patterns of dengue incidence, geographical features, outlier features, and selected climatic factors to create predictive features of dengue cases. Univariate correlation checks were then used to reduce dimensionality. For time series forecasting, SARIMA and SARIMAX models tuned by Auto-ARIMA were applied. These statistical models were compared to classical machine learning models obtained through TPOT, with hyperparameters tuned with Optuna. The Stochastic Gradient Descent (SGD) Regressor emerged as the best model, achieving a mean absolute error (MAE) of 32.26, a root mean squared error (RMSE) of 59.32, and an R-squared (R^2) value of 83.40%. This demonstrated significant predictive accuracy compared to other models. However, the model did not show potential in accurately predicting dengue outbreaks, indicating that more sophisticated features are needed for precise outbreak predictions.**

***Index Terms*—Dengue, Time Series Analysis, Predictive Modeling, Machine Learning, Climate Data, Philippines**

I. Introduction

Globally, dengue fever continues to pose a major public health concern. The disease stems from the dengue virus, transmitted to humans through the bites of infected Aedes mosquitoes, with Aedes aegypti being the primary vector [29]. The infection spectrum of dengue is wide, ranging from asymptomatic or sub-clinical infections to symptomatic ones, where it can manifest as mild, typical dengue or progress to severe forms such as dengue hemorrhagic fever (DHF) and dengue shock syndrome (DSS), both of which increase the risk of mortality [11]. By April 30, 2024, the World Health Organization has been notified of over 7.6 million dengue cases, with 3.4 million confirmed cases, more than 16,000 severe cases, and over 3,000 fatalities [30]. These numbers highlight a significant global increase in dengue cases over the past five years, straining healthcare systems and incurring economic loss due to direct medical expenses, non-medical costs, and productivity losses [23].

The dengue situation in the Philippines is particularly concerning. Annually, the Department of Health (DOH)-Epidemiology Bureau records thousands of cases and deaths, recently peaking at 23.5K deaths last 2019 [1] yet there is an alarming declining trend of dengue vaccine confidence in the country plummeting from 93% in 2015 [12] to a mere 33% in 2021 [31]. Recently, news report had already raised the alarming cases of dengue in the country ahead the onset of La Niña. Data from the DOH showed that there are already 59,267 dengue cases reported from January 1 to May 4, 2024 nationwide. This is higher compared to the 45,722 dengue cases during the same period in 2023 [10].

In the recent 6th Asia Dengue Summit held in June 2023 in Thailand, attended by 51 speakers and 451 delegates from over 24 countries including Philippines, a plan themed "Road Map to Zero Dengue Deaths" was formulated emphasizing the importance of various policies combining continued multisectoral collaboration (health, policy, education, environment), international mobilization of resources (research, funding, technology), sustained financial and political commitment, strengthened public health capacity, ongoing vector control efforts, innovations in dengue diagnostics and therapeutics, and vaccine advocacy to achieve the target of zero dengue deaths [25], hence the intention of this study to be part of this collaborative effort.

A. Objectives

The research aims to develop a robust predictive model for dengue cases within a specific cluster of locations in the Philippines. This clustered approach ensures that regional variations are accounted for. With machine learning techniques and time series analysis, models will be trained on the climate data and historical patterns of dengue incidence and will be used to predict cases in locations that belong to the same cluster. The predictions generated will then be utilized to identify potential dates for an outbreak. By identifying locations at heightened risk, timely interventions and strategic resource allocation could be facilitated, establishing appropriate guidelines that will help in the fight against dengue in the Philippines.

B. Contributions of the Study

This study introduces an approach for predicting dengue cases by leveraging climate variables in a machine learning model trained on time-series data from clustred multiple locations, grouped based on climatic variations. The method extends the conventional approach of creating separate time-series predictive models for each specific location by developing a partially generalized model. By simplifying

the consideration of spatial dependencies among locations, this approach eliminates the necessity for deep learning models, which are typically used in spatial-temporal analysis. Furthermore, this study encompasses a broader range of locations in the Philippines compared to existing dengue case prediction studies based on time-series data available at the time of this research.

II. Related Research

Understanding the interactions among humans, vectors of dengue fever, and environmental systems is crucial for controlling the complex dynamics of dengue transmission [21]. Climate factors, in particular, significantly influence the epidemiology of this disease [2]. Studies conducted across different regions have highlighted this influence.

Francisco et al. observed that precipitation patterns and landscape features significantly affect Aedes mosquitoes in Metro Manila (2012-2014) [9]. Faruk et al. found that precipitation, humidity, and air pressure were key factors in dengue transmission in Sri Lanka (2015-2019), with wetter zones being more susceptible to outbreaks [8]. In Khon Kaen, Thailand (2006-2016), Phanitchat et al. discovered that dengue outbreaks coincided with the rainy season, and there was a positive correlation between maximum temperatures and dengue incidence [18]. In Malaysia (2011-2019), Sarbhan Singh et al. found positive correlations between rainfall, temperature, and dengue incidence, with wind speed showing a negative correlation [24]. In Singapore (2012-2022), Na Tian et al. demonstrated significant associations between dengue incidence and factors such as UV index, solar radiation, and solar energy. They also showed that cloud cover is a key predictor of dengue incidence [26]. These findings highlight the significant impact of climatic conditions on dengue in tropical and subtropical regions.

Given the importance of climatic factors, advanced statistical models and machine learning algorithms were leveraged to predict dengue incidence rates effectively [14]. Navarro Valencia et al. used SARIMA, SARIMAX, and RNN-LSTM models to study dengue in Panama City (1999-2017), with the RNN-LSTM model performing best [28]. Nurul Azam Mohd Salim et al. found that the Support Vector Machine (SVM) was the most accurate for predicting dengue outbreaks in Selangor, Malaysia (2013-2017) [17]. Kakarla et al. analyzed dengue in Kerala, India (2003-2017), finding the LSTM model performed best [22].

In the Philippines, studies have used these techniques to predict dengue incidence. Mendoza developed an ARIMA model with an R^2 value of 0.739 for predicting dengue incidence in Magalang, Pampanga (2016-2019), discovering that rainfall was the most significant predictor [15]. Addawe et al. also incorporated climatic and time-related factors and compared the performance of machine learning models for predicting dengue cases in local communities in Baguio City (2016-2020). They found that Linear Regression performed best with an R^2 value of 0.5654 followed by Random Forrest with an R^2 value of 0.3778 [3]. Carvajal et al. used statistical and machine learning algorithms to predict dengue incidence in Metropolitan Manila (2009-2013), discovering that Random Forest with lagged climatic factors provided the lowest RMSE of 0.21 (considering log transformed dengue incidences) [6]. Buebos-Esteve and Dagamac integrated remote sensing and machine learning for predicting dengue incidence in the Philippines (2016-2020) where a Random Forest model performing best with an RMSE of 49.14 [5].

III. Methodology

Figure 1 presents the overview of the main steps undertaken in this study. In the following sections, details of each step is discussed providing further information of the dataset and how it was preprocessed, feature processing techniques, methods done to select the appropriate machine learning models and the tests done to evaluate the best model.

A. Data Collection Process

The dataset used in the study are composed of two separate data, the dengue cases and the climate variables, which were combined to create the initial set of input features and target.

1) Dengue Cases: A readily available public use information of confirmed dengue cases and deaths reported weekly from January 10, 2016 to January 10, 2021 covering 126 various locations all around the Philippines. The report was provided by the Department of Health-Epidemiology Bureau in the Philippines and hosted by the Humanitarian Data Exchanged[1] labeled as `'DOH-Epi Dengue Data 2016-2021'`. As noted on the webpage, reported dengue cases includes suspect, probable and confirmed, while deaths were also included in the total number of cases.

2) Climate Variables: Obtained from the National Aeronautics and Space Administration (NASA) Langley Research Center (LaRC) Prediction of Worldwide Energy Resource (POWER) Project [2]. As the project outlined, the satellite and model-based products have been shown to be sufficiently accurate to provide reliable solar resource data at $1x1$ degree resolution and meteorological quantities at $0.5x0.625$ degree resolution over regions where surface measurements are sparse or nonexistent. It offers two unique features: global data, and data that is generally contiguous in time, which are important characteristics to generate very large data archives.

The project provides a restful Application Programming Interfaces (API) support analysis ready data distribution service which enabled the researchers to obtain bulk of the variables for each specific location using its coordinates. Configurations was performed to select which climate feature variables, further discussed on Section III.C and presented in Table X from each group of available parameters will be transmitted by the API. Additionally, 'daily' temporal settings from the avaiable time intervals('climatology', 'monthly', 'daily', and 'hourly') was chosen.

B. Data Preprocessing

Temporal Alignment. To align the two time series data sets with different temporal granularities (weekly for dengue cases and daily for climate variables), the daily climate data was aggregated into weekly values. This approach was chosen as decomposing the weekly dengue

[1]https://data.humdata.org/dataset/philippine-dengue-cases-and-deaths
[2]https://power.larc.nasa.gov/

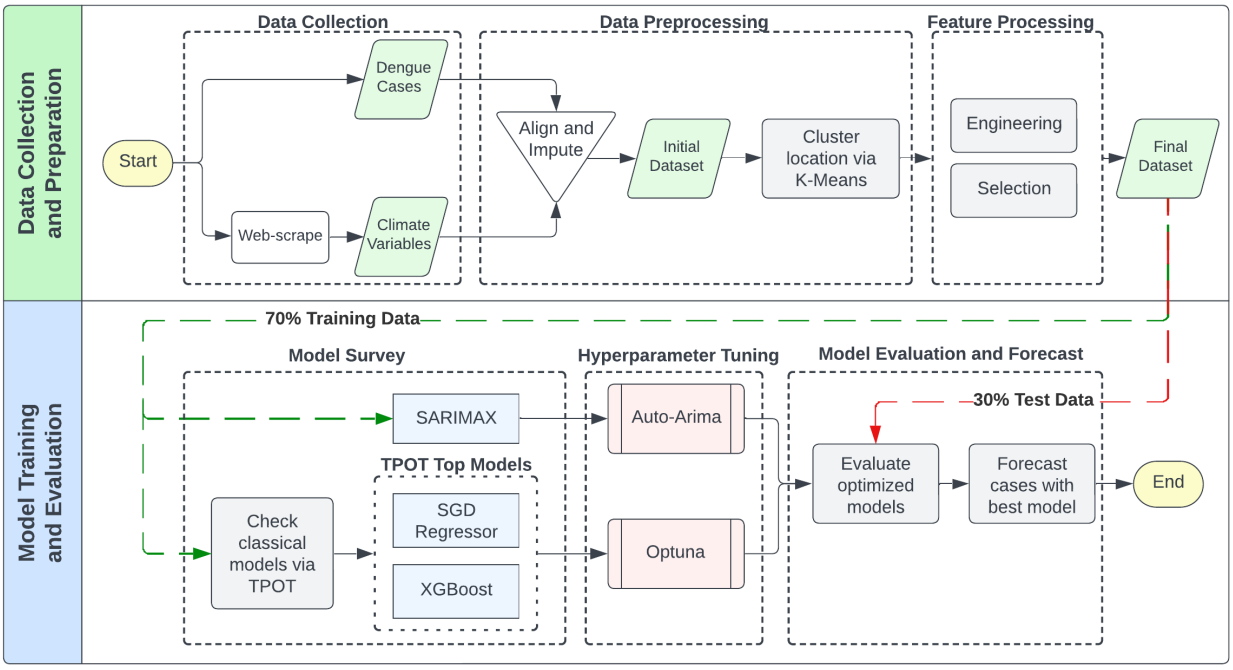


Fig. 1: Workflow diagram of the study

case data into daily values could lead to more inaccurate estimates and would generally be more challenging to perform. For aggregation, options such as weekly sum, mean, maximum value, and/or minimum value were considered. Researchers selected the most appropriate option for each climate variable, such as using the weekly sum for total precipitation or selecting the maximum or minimum value for extreme values like maximum temperature. Aggregation method used for each features is also presented in Table X.

Data Imputation. It was also discovered that some weeks were missing data for dengue cases. To ensure data completeness, these gaps were filled using the forward fill method, where the last known value is carried forward to fill missing entries.

Location Clustering. After ensuring that the dataset was aligned and complete, clustering was performed to group locations, ensuring that climatic variations were taken into account. Additionally, this step helped narrow down the scope for researchers by focusing on one specific clusters.

The clustering was achieved by combining t-SNE Dimensionality Reduction and K-Means Clustering in a single pipeline, optimized with Optuna by maximizing the silhouette score. For context, silhouette score is used to analyze the separation distance between the resulting clusters. It is especially useful if there is no prior knowledge of what is the true label for each object, which is the case for this study. For a pair of clusters A and B, the mean silhouette score s_i is computed as:

$$s(i) = \frac{b(i) - a(i)}{\max(b(i), a(i))}.$$

where $i \in A$, and $a(i)$ is the mean distance associated with point i to all the other points in cluster A. Similarly, $b(i)$ is the mean distance associated with point i to all the points of cluster B. Higher silhouette score indicates better quality of the clusters.

The range of the parameter values optimized for the pipeline is presented in Table I.

TABLE I: Hyperparameter Range for t-SNE and K-means

Algorithm	Hyperparameter	Range
t-SNE	number of components	2 - 3
	perplexity	5 - 50
	learning rate	10 - 1000
K-means	number of clusters	20 - 40
	initialization method (init)	k-means++, random
	initialization runs (n_init)	1-30

Stationarity Check. To determine if the data is stable and predictable, the mean and variance should remain constant over time, indicating that the time series is stationary which is crucial for time series models.

For this study, the Augmented Dickey-Fuller (ADF) Test [16], one of the most commonly used statistical tests for analyzing the stationarity of a series was employed and performed on individual time series data of dengue cases and climate variables.

The ADF test is a statistical significance test involving hypothesis testing with a null and alternate hypothesis. It computes a test statistic and reports p-values. Additionally, it also belongs to the category of tests called *'Unit Root Tests'* wherein the null hypothesis implies the presence of a unit root. If not rejected, the series is considered non-stationary. Hence apart from the p-value, the test statistics could also be checked such that if it is lower than the critical value, the null hypothesis is rejected, and the time series is inferred to be stationary.

C. Feature Processing

Several steps were taken to create and determine relevant features using the merged dataset of dengue cases and climate variables, making a dataset suitable for predictive analysis of dengue cases in Philippine locations.

Temporal Feature Engineering. Temporal features were derived by converting the date field into a DateTime type. Subsequently, the day, month, and year were extracted and ordinally encoded, allowing the incorporation of temporal trends in the model.

To guide the model in understanding the patterns and relationship of dengue cases over time, lagged average of values of cases calculated over a 4-week window preceding the prediction week was considered as a feature, this is computed as,

$$\text{Lagged Moving Average}_t = \frac{1}{n} \sum_{i=t-n-1}^{t-1} \text{cases}_i$$

where $n \in \{0, 1, 2, 3\}$.

To enhance the model's ability to recognize seasonality prominent in the dengue case dataset established in Section IV.A, the sine and cosine of the Fourier series for the weeks of the year were computed. The Fourier series transforms the time series data into components representing periodic patterns. This periodic transformation can be mathematically expressed as:

$$\text{cases\_fourier\_sin}_k(w) = \sin\left(2\pi k \frac{w}{p}\right)$$

$$\text{cases\_fourier\_cos}_k(w) = \cos\left(2\pi k \frac{w}{p}\right)$$

where $k \in \{1, 2, 3\}$ represents the number of Fourier terms used, $p = 52$ represents the number of weeks in a year, and $w \in \{1, 2, .., 52\}$ represents the week number within a year. This transformation captures the cyclical nature of dengue case patterns over a year.

Geographical Feature Engineering. As different locations may experience different climatic conditions affecting dengue transmission, geographical features such as latitude and longitude were included to capture spatial variations. The corresponding location names were also label encoded to facilitate their use in the model.

Anomaly Detection. Anomaly detection was performed using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [3], an unsupervised machine learning algorithm that clusters data points based on density. The parameters set for DBSCAN were eps represented the maximum neighborhood distance between 2 data points, and min_samples represented the minimum number of data points required in a neighborhood for a point to be considered a core point. These values were optimized via grid search, maximizing the silhouette score for a good clustering performance.

Climate Feature Selection. The climate variables collected and discussed previously were manually selected based on their potential impact on predicting dengue incidence. As shown in Table X, these features can be divided based on their relevance to environmental factors–solar radiation (Group 1), cloud cover (Group 2), hydrological variables (Group 3), temperature (Group 4), and wind speed (Group 5),

Solar radiation and ultraviolet (UV) index are critical predictors because they can accelerate mosquito development and increase breeding rates [26]. Cloud cover influences the local microclimate including temperature and humidity, impacting mosquito life cycles and breeding conditions [26]. Hydrological variables, including measures of soil moisture, precipitation, and humidity, provide insights into the availability of standing water, either creating abundant breeding sites or flooding out breeding sites [8] [9] [18] [24]. Temperature variables capture the thermal conditions experienced by mosquitoes, influencing their population biology and extrinsic incubation period [18] [24]. Wind speed, the final category, suppresses the flying activity of mosquitoes, influencing their contact with humans and potential dengue transmission [24]. These categorizations are supported by existing studies, some of which were discussed in Section II.

Univariate Correlation Check. A univariate correlation analysis was performed to address multicollinearity among the normalized features. This analysis identifies highly correlated features and retains only one of them. Initially, a correlation matrix of the normalized dataset was computed based on the Pearson method. A threshold of 0.95 was set for the absolute values of the correlation coefficients, where feature pairs with a correlation above this threshold were deemed highly correlated, excluding self-correlation. For each identified pair, one feature was selected for removal, effectively reducing the dimensionality of the dataset. This process of removing highly correlated features helps avoid redundancy and overfitting and improves model interpretability.

Temporal Splitting and Time Series Cross-Validation. Temporal splitting involves dividing the dataset into training and testing sets based on time, where the training set contains data up to a certain date and the testing set contains data after a certain date. For this study, the dataset was sorted by date in chronological order and a temporal split was performed. To be specific, 70% of the data, ranging from January 10, 2016, to July 14, 2019, was designated for training. The remaining 30%, ranging from July 21, 2019, to January 10, 2021, was designated for testing.

Traditional cross-validation techniques that randomly shuffle data do not suit time series data due to the data's inherent dependencies on temporal sequences. Instead, Time Series Cross-Validation (CV) [4] with five splits was employed on the training set derived from the temporal splitting of this study. This method maintains the temporal order of the data in each split, progressively expanding the training set for every fold while testing was done on subsequent unseen data. Time Series CV was performed during tuning with tools such as TPOT and Optuna, where after tuning, the model's performance was evaluated on the testing set from the temporal splitting.

D. Model Survey

Statistical Time Series Models. Given the nature of the problem, the standard statistical time series model, Autoregressive Integrated Moving Average (ARIMA), was used in

[3]https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN

[4]https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit

this study. ARIMA is a method first introduced by Box and Jenkins [4] and has now become one of the most popular methods for time series forecasting. However, given the prominent seasonality in the dataset, the variation of the classical ARIMA model, namely the seasonal ARIMA (SARIMA) model, is used. The seasonal ARIMA model is generally referred to as SARIMA (p,d,q)x(P,D,Q)$_s$, where p, d, q, and P, D, Q are non-negative integers that refer to the polynomial order of the autoregressive (AR), integrated (I), and moving average (MA) parts of the non-seasonal and seasonal components of the model, respectively.

The SARIMA model is described mathematically as follows:

$$\varphi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D y_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t$$

where:

- y_t is the forecast variable (i.e., PV production)
- $\varphi_p(B)$ is the regular AR polynomial of order p
- $\theta_q(B)$ is the regular MA polynomial of order q
- $\Phi_P(B^s)$ is the seasonal AR polynomial of order P
- $\Theta_Q(B^s)$ is the seasonal MA polynomial of order Q

The differentiating operator ∇^d and the seasonal differentiating operator ∇_s^D eliminate the non-seasonal and seasonal non-stationarity, respectively. B is the backshift operator, which operates on the observation y_t by shifting it one point in time (i.e., $B^k(y_t) = y_{t-k}$). The term ε_t follows a white noise process, and s defines the seasonal period. The polynomials and all operators are defined mathematically as follows:

$$\varphi_p(B) = 1 - \sum_{i=1}^{p}\varphi_i B^i \qquad \Phi_P(B^s) = 1 - \sum_{i=1}^{p}\Phi_i B^{s,i}$$

$$\theta_q(B) = 1 - \sum_{i=1}^{q}\theta_i B^i \qquad \Theta_Q(B^s) = 1 - \sum_{i=1}^{Q}\Theta_i B^{s,i}$$

$$\nabla^d = (1-B)^d \qquad \nabla_s^D = (1-B^s)^D$$

Extending this further, the statistical model version which considers the exogenous variables, SARIMAX, appropriate for the multivariate time series dataset was also studied. SARIMAX considers the exogenous variables as covariates and looks at them as external elements that influence the time series under study. It captures both short-term and long-term dependencies within the data, making it a robust tool for forecasting. The SARIMAX is generally expressed as follows:

$$\varphi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D y_t = \beta_k x_{k,t}{}' + \theta_q(B)\Theta_Q(B^s)\varepsilon_t$$

where $x_{k,t}$ is the vector including the kth explanatory input variables at time t and β_k is the coefficient value of the kth exogenous input variable. The stationarity and invertibility conditions are equal to those of ARMA models [27].

Classical Machine Learning Models. In addressing time series problems, a diverse array of classical machine learning models can be applied. To help narrow down the options, the Python Automated Machine Learning (autoML) tool, TPOT (Tree-based Pipeline Optimization Tool)[5] [13] was employed. TPOT optimizes pipelines using genetic programming and explores thousands of possible pipelines to find the best one for the given dataset. From the results of the autoML, the researchers selected the best performing model(s) then integrated it with the established methodology of this study.

E. Hyperparameter Optimization

1) Auto-Arima: For statistical time series models, tuning of their parameters was performed with auto_arima [6], which works to find the optimal order of parameters by minimizing the Akaike Information Criterion (AIC). It is designed to perform a grid search over different combinations of defined ranges presented in Table II for p, d, q, and P, D, Q values to find the best fit for the data.

TABLE II: Hyperparameter Ranges for SARIMA and SARIMAX

Hyperparameter	Range
p	0 to 5
d	0 to 2
q	0 to 5
P	0 to 5
D	0 to 2
Q	0 to 5

2) Optuna: For tuning the hyperparameters of the classical models, Optuna [7] was employed. Optuna is an open-source optimization library designed to fine-tune the hyperparameters of machine learning models. It surpasses traditional methods like grid and random search by employing more sophisticated algorithms such as the Tree-structured Parzen Estimator (TPE), allowing for a more effective and faster search process. This required an objective function specifying the model and the hyperparameters to tune, while using a validation metric to maximize or minimize. For this study, the range for the values of the hyperparameters was iteratively refined by continuously running the optimization process and checking the slice plot to determine the range of values where the minimum RMSE, which was designated as the validation metric to minimize, is likely to occur. Based on the top-performing models identified in the initial survey, the range for hyperparameters for SGD and XGBoost Regressor are presented in Table III.

TABLE III: Hyperparameter Range for SGD and XGBoost Regressor

Algorithm	Hyperparameter	Range
XGBoost	n estimators	50 to 200
	max depth	2 to 32 (log scale)
	learning rate	0.01 to 0.1 (log scale)
	subsample	0.1 to 0.7
	min child weight	8 to 400
SGD Regressor	alpha	1e-6 to 1e-1
	eta0	1e-4 to 1e-1
	fit intercept	True, False
	l1 ratio	0.0 to 1.0
	learning rate	constant, optimal, invscaling, adaptive
	loss	epsilon insensitive, huber, squared error, squared epsilon insensitive
	penalty	l2, l1, elasticnet
	power t	0.1 to 1.0

[5]http://epistasislab.github.io/tpot/

[6]https://alkaline-ml.com/pmdarima/modules/generated/pmdarima.arima.auto_arima.html

[7]https://optuna.org/

F. Model Evaluation

The forecasting algorithm is applied to all the optimized models, and the results are analyzed to assess their performance. The analysis includes the computation of forecast error statistics based from the actual test dataset and the comparison of results across the candidate models.

Mean absolute error (MAE). One of the simplest and most intuitive evaluation metrics for a time series forecasting model. It measures the average magnitude of the absolute errors between the actual and predicted values, without considering the direction of the errors. MAE is calculated as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_{\text{true},i} - y_{\text{pred},i}|$$

where n is the number of observations, y_{true} is the actual value, and y_{pred} is the predicted value. MAE is easy to interpret, as it tells how much, on average, predictions deviate from the actual values. However, MAE is not sensitive to outliers or large errors, and it does not account for the scale of the data.

Root mean squared error (RMSE). Another common evaluation metric for a time series forecasting model. It measures the average magnitude of the squared errors between the actual and predicted values, and then takes the square root. RMSE is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{\text{true},i} - y_{\text{pred},i})^2}$$

where n is the number of observations, y_{true} is the actual value, and y_{pred} is the predicted value. RMSE is similar to MAE, but it gives more weight to larger errors, as they are squared before averaging. RMSE is also more sensitive to outliers and variations in the data. However, RMSE is not easy to interpret, as it does not have the same unit as the original data.

R-squared (R^2). Measures the proportion of variance in the dependent variable that is explained by the independent variables in the model, particularly insightful for multivariate time series problems. It is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_{\text{true},i} - y_{\text{pred},i})^2}{\sum_{i=1}^{n}(y_{\text{true},i} - \bar{y}_{\text{true}})^2}$$

where $\bar{y}_{\text{true}}$ is the mean of the actual values. It ranges from 0 to 1, where 1 indicates a perfect fit and 0 indicates no relationship between variables.

The important point to note in the case of R^2 is that it does not show if the model is satisfactory for future predictions or not. It shows if the model is a good fit for the observed values and how good of a "fit" it is. High R^2 means that the correlation between observed and predicted values is high, not how well it was able to predict future values.

G. Outbreak Prediction

The seasonal increase in dengue cases, typically observed during or just after the rainy season, must be distinguished from an unexpected surge in cases beyond a defined threshold, commonly referred to as an outbreak. This should not be confused with the number of reported cases exceeding certain expected levels, which is more appropriately termed "aberrations" [7]. Classifying a dengue outbreak as an "unexpected increase in cases" requires additional criteria and efforts.

Various methods exist to define an outbreak, as detailed in the WHO technical handbook for dengue surveillance, dengue outbreak prediction/detection, and outbreak response [20]. One such method is the "endemic channel," which delineates security, alarm, and outbreak zones based on expected case levels using the weekly (or monthly) average number of cases over the last five years. However, this approach has significant disadvantages, such as previous outbreaks resulting in thresholds that are too high, the lack of a satisfactory algorithm for recognizing past aberrations, and the inclusion of outbreak years in the calculation of the historical (moving) average, which can create excessively high thresholds for subsequent years.

For this reason, the "moving average" method [19] is used in this study, as it is suitable for diseases with moderate or high prevalence and seasonal fluctuations. It provides a timely, specific, and statistically based signal for control efforts. The threshold is computed as the sum of the moving average of three 4-week dengue case periods plus the value of two standard deviations above the number of dengue cases for the cases four weeks prior.

Let cases(t) be the number of dengue cases in week t. The 4-week moving average ending in week t is defined as:

$$MA_4(t) = \frac{1}{4}\sum_{i=0}^{3}\text{cases}(t-i).$$

The moving average of three consecutive 4-week periods ending in week t is:

$$MA_{3\times4}(t) = \frac{1}{3}\sum_{j=0}^{2}MA_4(t-4j).$$

The standard deviation of the dengue cases over the 4 weeks prior to week t is computed as:

$$SD_4(t) = \sqrt{\frac{1}{4}\sum_{i=0}^{3}(\text{cases}(t-i) - MA_4(t))^2}.$$

Finally, the outbreak threshold for week t is:

$$\text{Threshold}(t) = MA_{3\times4}(t) + 2 \times SD_4(t).$$

An outbreak is determined if the number of cases in week t exceeds the outbreak threshold:

$$\text{Outbreak}(t) = \begin{cases} 1 & \text{if cases}(t) > \text{Threshold}(t) \\ 0 & \text{otherwise} \end{cases}$$

For this study, this method was employed to predict the possible outbreaks for the year 2019-2020 for Bulacan, Rizal, and Quezon City by using the weekly predicted dengue cases.

IV. Results and Discussion

This section is composed of two sub-sections, similarly presented by Figure 1, which are data preparation and model training and evaluation. Here, the results and insights gained from all the steps discussed in Section III will be presented in detail.

A. Data Preparation

After integrating the dengue cases and climate variables, the initial dataset resulted in 33,012 datapoints with 54 climate features and one target variable. This dataset was then clustered using the t-SNE + K-Means pipeline, resulting in 28 location clusters with an average silhouette score of 0.42.

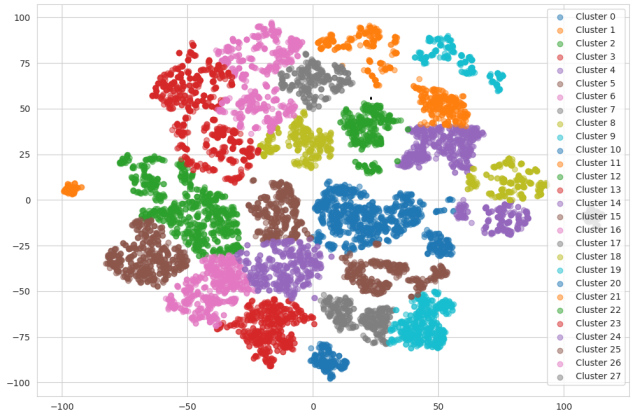


Fig. 2: Clustering of locations based on the dengue cases and complete set of climate variables

Details of the tuned hyperparameters after performing Optuna optimization for the dimensionality reduction-clustering algorithm are presented in Table IV, while the scatter plot of the clusters is shown in Figure 2. From the plot, it could be observed that although there are no overlapping points, some data points are very close to each other considering that some locations were actually considered in multiple clusters. Intuitively, most of the clusters contain locations that are geographically closer to each other, emphasizing their similarity in climatic factors.

Arbitrarily, from the set of clusters, the researchers chose one cluster containing **Bulacan, Quezon City, and Rizal** to focus on.

TABLE IV: Tuned Hyperparameters of t-SNE and K-Means Clustering

Algorithm	**Hyperparameter**	**Range**
t-SNE	number of components	2
	perplexity	39.14
	learning rate	527.07
K-means	number of clusters	28
	initialization method (init)	k-means++
	initialization runs (n_init)	23

Focusing on the three locations, the initial large dataset was reduced to 783 timeseries datapoints. Performing the ADF test on the individual time series of dengue cases and climate variables for the three locations revealed their stationarity, suggesting stability and potential predictability. The results of the ADF test for dengue cases are presented in Table V. It can be observed that the test statistics for all locations are lower than the critical value at 5%, leading to the rejection of the null hypothesis of a unit root, indicating stationarity. This is further supported by the small p-values, which are significantly less than the 5% significance level, also resulting in the rejection of the null hypothesis.

Furthermore, decomposing the dengue cases helped in understanding the inherent behavior of the data. The plot

TABLE V: ADF Test Result for the Dengue Cases

Location	Test Statistics	Critical Value (5%)	p-Value
Bulacan	-4.6872	-2.8729	8.89e-5
Rizal	-3.9619	-2.8729	0.0016
Quezon City	-3.8470	-2.8730	0.0025

presented in Figure 3 highlights the recurring and predictable patterns in the seasonality plot, emphasizing the prominent seasonality of the data. This provides a strong motivation to consider seasonal variation in the study, which necessitates the use of seasonal variation of ARIMA and the consideration of Fourier series-related features as discussed in Section III.C.

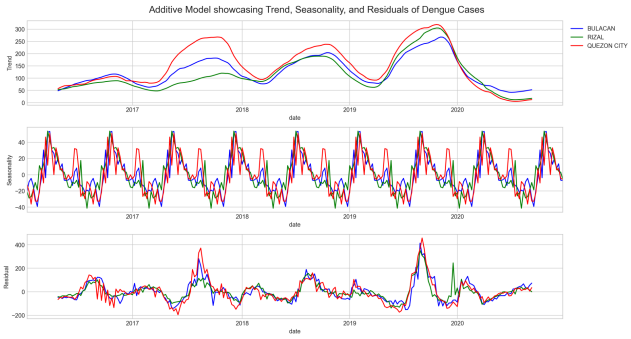


Fig. 3: Decomposed dengue cases time series data

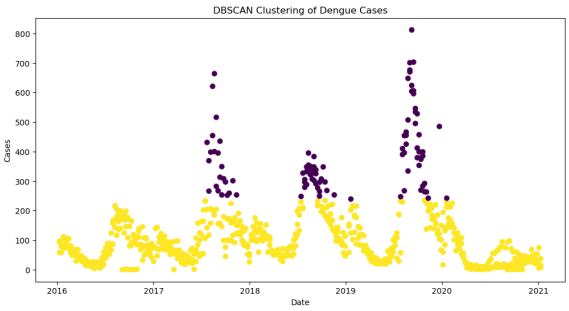


Fig. 4: Anomaly detection of dengue cases

TABLE VI: Tuned Hyperparameters of Anomaly Detection with DBSCAN

Hyperparameter	**Range**
epsilon	10
minimum number of samples	20

Subsequently, the steps presented in Section III.C were performed to add meaningful features to the data. The first step was anomaly detection of the dengue cases using DBScan, which obtained a silhouette score of 0.72 and identified 14% of the dengue cases as outliers, primarily composed of points belonging to weeks with increasing spikes as seen on the scatter plot of the results in Figure 4. The tuned hyperparameters are presented in Table VI.

With the completion of feature engineering, the feature selection was then performed which successfully filtered out 19 climate features resulting to the 40 features shown in the correlation plot at Figure 5.

B. Model Training and Evaluation

TPOT Top Performing Models. Table VII shows a survey of classical model performance results from TPOT

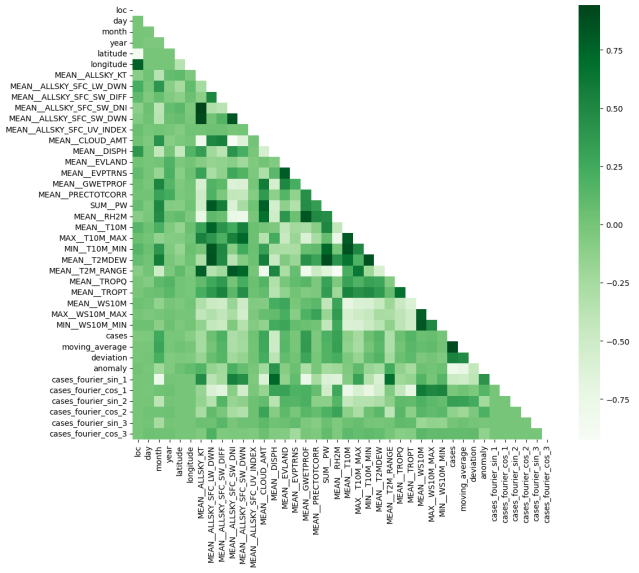


Fig. 5: Correlation matrix of the remaining features after thresholding

applied to predicting dengue cases for Bulacan, Quezon City, and Rizal. The table lists the types of regression models evaluated, details the hyperparameters used for each model, and the performance metric of each model. Here, the scores in negative thousands represent the negative mean squared error (MSE) values, indicating that models with less negative scores perform better. The top two models in this table were the Stochastic Gradient Descent method for regression (SGD Regressor) and the Extreme Gradient Boosting for regression (XGBoost Regressor) with the corresponding scores of -3340.9179 and -3732.0512. The SGD Regressor performed best due to its use of the Huber loss function, which is robust to outliers, and ElasticNet regularization. On the other hand, XGBoost Regressor was already known for its ability to handle complex relationships in the data. To get better model results, hyperparameter tuning was done on these models with Optuna.

Hyperparameter Tuning. Although TPOT already provided an initial optimization of hyperparameters, additional tuning with Optuna is necessary as the tool explores a broader and finer hyperparameter space focused on a model, evaluating more combinations of hyperparameter values for a more optimized performance. Table VIII shows the hyperparameters tuned using AutoArima and Optuna, where the models tuned with AutoArima were SARIMA and SARIMAX while the models tuned with Optuna were XGBoost Regressor and SGD Regressor, which were the top 2 models evaluated by TPOT. It is worth noting that in the table, SARIMAX offers a simpler configuration for seasonal and non-seasonal components compared to SARIMA. This simplicity is due to the inclusion of exogenous factors that can help capture data patterns, reducing the need for a more complex statistical model.

Model Evaluation. The final performance metrics for the tuned models used in predicting dengue fever incidence are summarized in Table IX. The performance of these models was measured in terms of MAE, RMSE, and R^2 values. These metrics represent the average values across Bulacan, Quezon City, and Rizal.

The SARIMA model performed poorly, with an MAE of 130.87, an RMSE of 175.85, and a negative R^2 value of -5.14%. A negative R^2 value indicates that the model performed worse than a simple mean baseline prediction, likely due to its inability to incorporate external variables. In contrast, the SARIMAX model showed improvement, achieving an MAE of 77.37 and an RMSE of 89.36. The R^2 value of 73.14% highlights its strong predictive performance due to the inclusion of exogenous variables that influence dengue prediction such as the climatic factors used in this study. XGBRegressor, an implementation of the gradient-boosting algorithm, demonstrated an MAE of 58.48, an RMSE of 91.08 and an R^2 value of 71.93%. Although its R^2 value is lower than SARIMAX, the model still showed good predictive capability due to its ability to handle complex nonlinear relationships between features. The SGDRegressor outperformed all other models, with the lowest MAE of 32.26, an RMSE of 59.32, and the highest R^2 value of 83.40%. Similar to the results obtained in TPOT, the SGDRegressor best captures the relationship between the variables for predicting dengue incidence.

From Figure 6, it is evident that SARIMA performed poorly, highlighting the need for necessary time series preprocessing. This study, however, intentionally skipped such preprocessing steps to focus on exploring the climate variables and feature engineered attributes. Additionally, it is noticeable in the time series plots that, unlike SGD Regressor and SARIMAX, which were able to predict possible spikes or increases in cases in the testing set (more pronounced in SGD), XGBoost has overfitted, indicating its poor performance in predicting sudden increases in the target variable.

Outbreak Prediction.The application of the established threshold for classifying outbreaks for the year 2019 to 2020 revealed several misclassifications, as illustrated in the Figure 7. Despite predicting one outbreak for Bulacan, the model predicted it before July when it actually occurred after that month. For Rizal, the model predicted 10 outbreaks compared to the actual 5. One instance was predicted in January 2020, but the outbreak occurred in the last quarter of 2019. However, some predictions coincided with actual outbreaks in early July 2020. Notably, the model failed to predict any outbreaks at the end of 2020. Similarly, for Quezon City, only one predicted outbreak before July 2019 matched the actual occurrence, which took place in August-September. The outbreaks in 2020 were also missed by the model.

These discrepancies highlight the sensitivity of the outbreak threshold and the significant impact of small deviations on outbreak classification. It also underscores the distinction between outbreaks and case spikes (or aberrations, as discussed in Section III.G), as some outbreaks exhibit gradual increases rather than sudden spikes. This complexity suggests more sophisticated features are actually needed to accurately predict outbreaks.

V. Conclusion and Recommendation

This study successfully developed predictive models for dengue cases in the Philippines, focusing on locations in

TABLE VII: Classical Model Survey Result of TPOT

Model	Hyperparameters	Score
SGD Regressor	alpha = 0.01, eta0 = 0.01, fit_intercept = False, l1_ratio = 0.25, learning_rate = invscaling, loss = huber, penalty = elasticnet, power_t = 0.5	-3340.9179
XGBoost Regressor	learning_rate = 0.1, max_depth = 6, min_child_weight = 17, n_estimators = 100, n_jobs = 1, objective = reg:squarederror, subsample = 0.5, verbosity = 0	-3732.0512
Linear SVR	C = 0.0001, dual=True, epsilon = 0.1, loss = squared_epsilon_insensitive, tol = 0.1	-3952.5170
Gradient Boosting Regressor	alpha = 0.8, learning_rate = 0.1, loss = huber, max_depth = 10, max_features = 0.5, min_samples_leaf = 17, min_samples_split = 14, n_estimators = 100, subsample = 0.9	-4268.3599
Random Forest Regressor	bootstrap = True, max_features = 0.4, min_samples_leaf = 16, min_samples_split = 14, n_estimators = 100	-4485.5252
AdaBoost Regressor	learning_rate = 0.1, loss = linear, n_estimators = 100	-4797.2179
ElasticNet CV	alpha = 0.019, l1_ratio = 0.65, tol = 0.1	-4816.1062
Decision Tree Regressor	max_depth = 5, min_samples_leaf = 8, min_samples_split = 17	-4876.1631
Ridge CV	alpha = 0.001, eta0 = 0.01, fit_intercept = True, l1_ratio = 0.0, learning_rate = constant, loss = huber, penalty = elasticnet, power_t = 0.5	-4893.4132
ExtraTrees Regressor	bootstrap = False, max_features = 0.35, min_samples_leaf = 8, min_samples_split = 18, n_estimators = 100	-4900.6924
KNeighbors Regressor	n_neighbors = 19, p = 2, weights = uniform	-5141.9216

TABLE VIII: Tuned Hyperparameters

Model	Hyperparameter
SARIMA	p = 2, d = 1, q = 3, P = 3, D = 0, Q = 2, m = 4
SARIMAX	p = 1, d = 0, q = 1, P = 0, D = 0, Q = 0, m = 4
XGBoost	'n_estimators' = 164, max_depth = 4, learning_rate = 0.0283, subsample = 0.452, min_child_weight = 14
SGDRegressor	'alpha' = 0.0304, fit_intercept = True, l1_ratio = 0.8563, loss = 'huber', penalty = 'elasticnet', power_t = 0.2232

TABLE IX: Model Performance Metrics

Model	MAE	RMSE	R-squared
SARIMA	130.87	175.85	-5.14%
SARIMAX	77.37	89.36	73.14%
XGBoost	58.48	91.08	71.93%
SGDRegressor	32.26	59.32	83.40%

the same cluster, namely, Bulacan, Quezon City, and Rizal. By integrating climate data and historical dengue incidence, advanced statistical and machine learning techniques managed to predict dengue cases in these locations. Among the models tuned and evaluated, the Stochastic Gradient Descent (SGD) Regressor had the best performance, achieving high predictive accuracy with a MAE of 32.26, RMSE of 59.32, and an R^2 value of 83.40%.

While promising results were obtained in predicting dengue cases, the models struggled to accurately predict outbreaks, highlighting the need for more sophisticated features and methods for outbreak prediction. Incorporating additional features beyond climate variables, such as population, demographics, access to healthcare, and city indices, may improve the accuracy of outbreak predictions. Additionally, extending the study to focus on outbreak classification, where the model predicts whether a specific location will experience an outbreak or not, could provide valuable insights.

In terms of improving the forecasting of dengue cases, various experimental setups could be explored to better understand the spatial dependencies of the cases. Training models on specific locations while using the cases from previous weeks of adjacent locations or similar clusters as features, or including all available locations with their coordinates as features, might help the model learn the spatial aspects of the problem. Finally, although this was a delimitation of the study, using deep learning and more complex models such as Long Short-Term Memory (LSTM) networks with attention mechanisms or Gated Recurrent Units (GRUs), which have been proven to perform well on similar datasets, could result in more robust models.

REFERENCES

[1] Philippine Dengue Cases and Deaths - Humanitarian Data Exchange. https://data.humdata.org/dataset/philippine-dengue-cases-and-deaths. [Accessed 12-06-2024].

[2] N. A. M. H. Abdullah, N. C. Dom, S. A. Salleh, H. Salim, and N. Precha. The association between dengue case and climate: A systematic review and meta-analysis. *One Health*, 15:100452, 2022.

[3] J. C. Addawe, J. D. Caro, and R. A. B. Juayong. Machine learning methods for modeling dengue incidence in local communities. In *Novel & Intelligent Digital Systems Conferences*, pages 392–400. Springer, 2022.

[4] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[5] D. E. Buebos-Esteve and N. H. A. Dagamac. Spatiotemporal models of dengue epidemiology in the philippines: Integrating remote sensing and interpretable machine learning. *Acta Tropica*, 255:107225, 2024.

[6] T. M. Carvajal, K. M. Viacrusis, L. F. T. Hernandez, H. T. Ho, D. M. Amalin, and K. Watanabe. Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan manila, philippines. *BMC infectious diseases*, 18:1–15, 2018.

[7] P. Farrington and N. Andrews. 203Outbreak Detection: Application to Infectious Disease Surveillance. In *Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance*. Oxford University Press, 12 2003.

[8] M. O. Faruk, S. Jannat, and M. S. Rahman. Impact of environmental factors on the spread of dengue fever in sri lanka. *International Journal of Environmental Science and Technology*, 19:10637 – 10648, 2022.

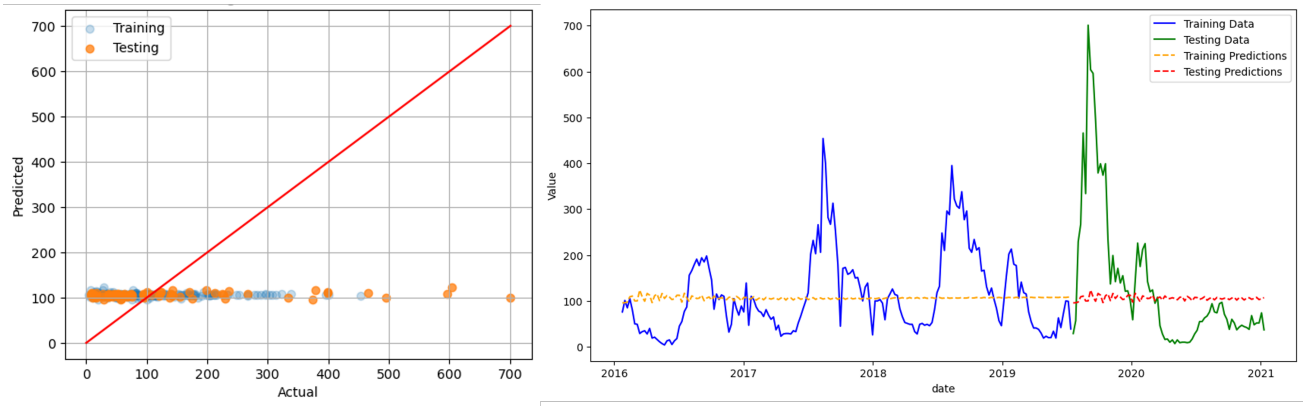
[9] M. E. Francisco, T. M. Carvajal, M. Ryo, K. Nukazawa, D. M. Amalin, and K. Watanabe. Dengue disease dynamics are modulated by the combined influences of precipitation and landscape: A machine learning approach. *Science of The Total Environment*, 792:148406, 2021.

[10] HDT. Philippines seeing more dengue cases in 2024, May 2024. Accessed: 2024-06-17.

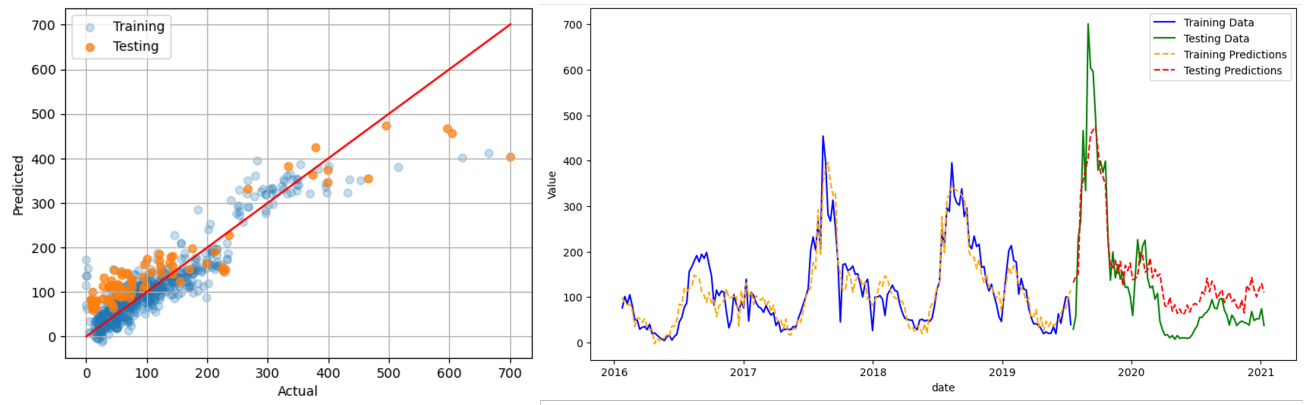
[11] Q. Jing and M. Wang. Dengue epidemiology. *Global Health Journal*, 3(2):37–45, 2019.

[12] H. Larson, K. Hartigan-Go, and A. de Figueiredo. Vaccine confidence plummets in the philippines following dengue vaccine scare: why it matters to pandemic preparedness. *Human Vaccines & Immunotherapeutics*, 15(3):625–627, 2019. [published correction appears in Hum Vaccin Immunother. 2020 Oct 2;16(10):2577. doi: 10.1080/21645515.2019.1628510].

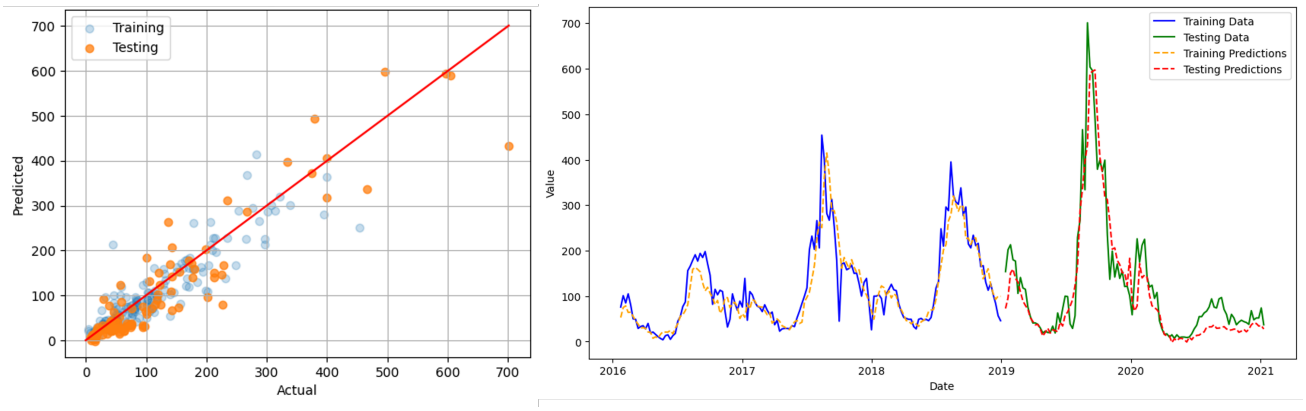
[13] T. T. Le, W. Fu, and J. H. Moore. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. *Bioinformatics*, 36(1):250–256, 2020.



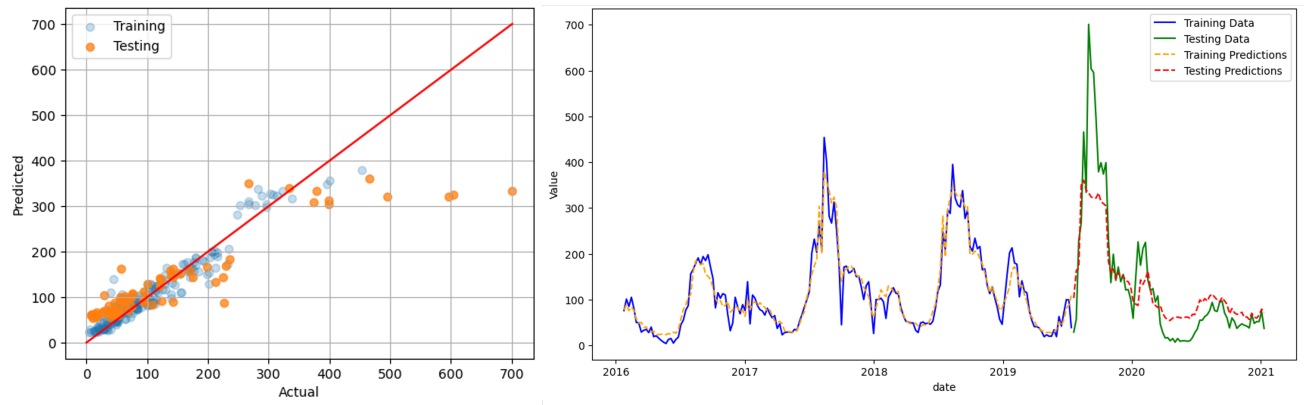
(a) SARIMA



(b) SARIMAX

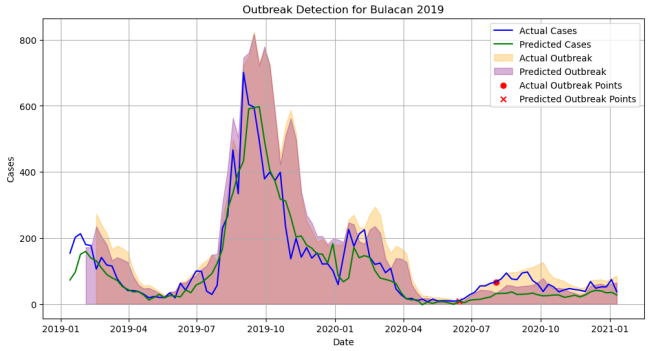


(c) SGD Regressor

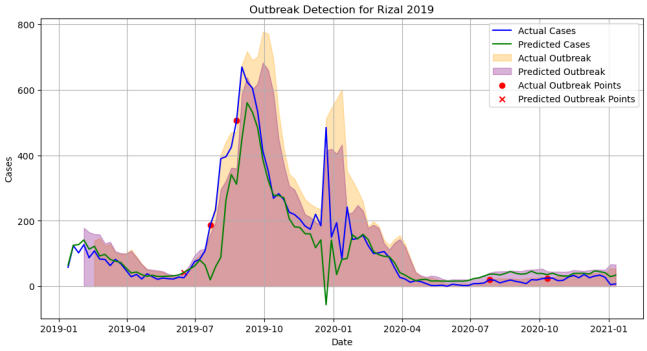


(d) XGBoost Regressor

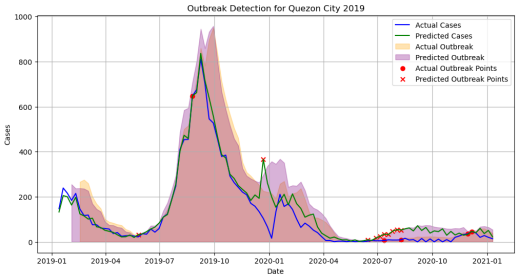
Fig. 6: Train-Test result of the evaluated models



(a) Bulacan



(b) Rizal



(c) Quezon City

Fig. 7: Actual vs predicted outbreaks for each location

[14] X. Y. Leung, R. M. Islam, M. A. Adhami, D. Ilić, L. McDonald, S. Palawaththa, B. Diug, S. U. Munshi, and M. N. Karim. A systematic review of dengue outbreak prediction models: Current scenario and future directions. *PLOS Neglected Tropical Diseases*, 17, 2022.

[15] A. P. Mendoza. Dengue incidence forecasting model in magalang, pampanga using time series analysis. *Informatics in Medicine Unlocked*, 44:101439, 2024.

[16] R. Mushtaq. Augmented dickey fuller test. 2011.

[17] C. R. Nurul Azam Mohd Salim, Yap Bee Wah et al. Prediction of dengue outbreak in selangor malaysia using machine learning techniques. *Scientific Reports*, 11:939, 2021.

[18] T. Phanitchat, B. Zhao, U. Haque, C. Pientong, T. Ekalaksananan, S. Aromseree, K. Thaewnongiew, B. Fustec, M. J. Bangs, N. Alexander, and H. J. Overgaard. Spatial and temporal patterns of dengue incidence in northeastern thailand 2006–2016. *BMC Infectious Diseases*, 19, 2019.

[19] J. G. Rigau-Pérez, P. S. Millard, D. R. Walker, C. C. Deseda, and A. Casta-Vélez. A deviation bar chart for detecting dengue outbreaks in puerto rico. *American Journal of Public Health*, 89(3):374–378, 1999.

[20] S. Runge-Ranzinger, A. Kroeger, P. Olliaro, L. Bowman, O. Horstick, L. Lloyd, and P. McCall. *Technical handbook for dengue surveillance, dengue outbreak prediction/detection and outbreak response ("model contingency plan")*. 10 2016.

[21] M. Saputra and H. Oktaviannoor. One health approach to dengue haemorrhagic fever control in indonesia: A systematic review. *KnE Life Sciences*, 4:201–221, 2018.

[22] H. P. V. Satya Ganesh Kakarla, Phani Krishna Kondeti et al. Weather integrated multiple machine learning models for prediction of dengue prevalence in india. *International Journal of Biometeorology*, 67:285–297, 2023.

[23] D. S. Shepard, E. A. Undurraga, Y. A. Halasa, and J. D. Stanaway. The global economic burden of dengue: a systematic analysis. *The Lancet Infectious Diseases*, 16(8):935–941, 2016. PMID: 27091092.

[24] S. Singh, L. C. Herng, L. H. Sulaiman, S. F. Wong, J. Jelip, N. Mokhtar, Q. K. Harpham, G. Tsarouchi, and B. S. Gill. The effects of meteorological factors on dengue cases in malaysia. *International Journal of Environmental Research and Public Health*, 19, 2022.

[25] N. Srisawat, D. Gubler, T. Pangestu, U. Limothai, U. Thisyakorn, Z. Ismail, et al. Proceedings of the 6th asia dengue summit, june 2023. *PLoS Neglected Tropical Diseases*, 18(3):e0012060, 2024.

[26] N. Tian, J.-X. Zheng, L.-H. Li, J.-B. Xue, S. Xia, S. Lv, and X.-N. Zhou. Precision prediction for dengue fever in singapore: A machine learning approach incorporating meteorological data. *Tropical Medicine and Infectious Disease*, 9(4):72, 2024.

[27] S. I. Vagropoulos, G. I. Chouliaras, E. G. Kardakos, C. K. Simoglou, and A. G. Bakirtzis. Comparison of sarimax, sarima, modified sarima and ann-based models for short-term pv generation forecasting. In *2016 IEEE International Energy Conference (ENERGYCON)*, pages 1–6, 2016.

[28] V. N. Valencia, Y. Díaz, J. M. Pascale, M. F. Boni, and J. E. Sánchez-Galán. Assessing the effect of climate variables on the incidence of dengue cases in the metropolitan region of panama city. *International Journal of Environmental Research and Public Health*, 18, 2021.

[29] World Health Organization. *Dengue: Guidelines for Diagnosis, Treatment, Prevention, and Control*, 2009. https://www.who.int/publications/i/item/9789241547871.

[30] World Health Organization. Disease outbreak news: Dengue - global situation, May 2024. Accessed: 12 June 2024.

[31] V. G. Yu, G. Lasco, and C. C. David. Fear, mistrust, and vaccine hesitancy: Narratives of the dengue vaccine controversy in the philippines. *Vaccine*, 39(35):4964–4972, 2021.

ANNEX

TABLE X: Selected Climate Features from NASA POWER Project

Definition	Variable Name	Treatment	Unit	Description	Group
All Sky Insolation Clearness Index	ALLSKY KT	MEAN	-	A fraction representing clearness of the atmosphere; the all sky insolation that is transmitted through the atmosphere to strike the surface of the earth divided by the average of top of the atmosphere total solar irradiance incident.	1
All Sky Surface Longwave Downward Irradiance	ALLSKY SFC LW DWN	MEAN	MJ/m^2/day	The downward thermal infrared irradiance under all sky conditions reaching a horizontal plane the surface of the earth. Also known as Horizontal Infrared Radiation Intensity from Sky.	1
All Sky Surface Longwave Upward Irradiance	ALLSKY SFC LW UP	MEAN	MJ/m^2/day	The upward thermal infrared irradiance under all sky conditions.	1
All Sky Surface PAR Total	ALLSKY SFC PAR TOT	MEAN	MJ/m^2/day	The total Photosynthetically Active Radiation (PAR) incident on a horizontal plane at the surface of the earth under all sky conditions.	1
All Sky Surface Shortwave Diffuse Irradiance	ALLSKY SFC SWDIFF	MEAN	MJ/m^2/day	The diffuse (light energy scattered out of the direction of the sun) solar irradiance incident on a horizontal plane at the surface of the earth under all sky conditions.	1
All Sky Surface Shortwave Downward Direct Normal Irradiance	ALLSKY SFC SW DNI	MEAN	MJ/m^2/day	The direct solar irradiance incident to a horizontal plane normal (perpendicular) to the direction of the sun's position under all sky conditions.	1
All Sky Surface Shortwave Downward Irradiance	ALLSKY SFC SW DWN	MEAN	MJ/m^2/day	The total solar irradiance incident (direct plus diffuse) on a horizontal plane at the surface of the earth under all sky conditions. An alternative term for the total solar irradiance is the "Global Horizontal Irradiance" or GHI.	1
All Sky Surface Shortwave Upward Irradiance	ALLSKY SFC SW UP	MEAN	MJ/m^2/day	The upward shortwave irradiance under all sky conditions.	1
All Sky Surface UVA Irradiance	ALLSKY SFC UVA	MEAN	MJ/m^2/day	The ultraviolet A (UVA 315nm-400nm) irradiance under all sky conditions.	1
All Sky Surface UVB Irradiance	ALLSKY SFC UVB	MEAN	MJ/m^2/day	The ultraviolet B (UVB 280nm-315nm) irradiance under all sky conditions.	1
All Sky Surface UV Index	ALLSKY SFC UV INDEX	MEAN	-	The ultraviolet radiation exposure index	1

TABLE X: (continued)

Definition	Variable Name	Treatment	Unit	Description	Group
Cloud Amount	CLOUD AMT	MEAN	%	The average percent of cloud amount during the temporal period.	2
Cloud Amount at Daytime	CLOUD AMT DAY	MEAN	%	The average percent of cloud amount during daylight.	2
Cloud Amount at Nighttime	CLOUD AMT NIGHT	MEAN	%	The average percent of cloud amount during nighttime.	2
Zero Plane Displacement Height	DISPH	MEAN	m	The height at which the mean velocity is zero due to large obstacles such as buildings/canopy.	2
Evaporation Land	EVLAND	MEAN	kg m^{-2} s^{-1} 10^6	The evaporation over land at the surface of the earth.	3
Evapotranspiration Energy Flux	EVPTRNS	MEAN	MJ/m^2/day	The evapotranspiration energy flux at the surface of the earth.	3
Profile Soil Moisture	GWETPROF	MEAN	-	The percent of profile soil moisture a value of 0 indicates a completely water-free soil and a value of 1 indicates a completely saturated soil; where profile is the layer from the surface down to the bedrock.	3
Root Zone Soil Wetness	GWETROOT	MEAN	-	The percent of root zone soil wetness a value of 0 indicates a completely water-free soil and a value of 1 indicates a completely saturated soil; where root zone is the layer from the surface 0 cm to 100 cm below grade.	3
Surface Soil Wetness	GWETTOP	MEAN	-	The percent of soil moisture a value of 0 indicates a completely water-free soil and a value of 1 indicates a completely saturated soil; where surface is the layer from the surface 0 cm to 5 cm below grade.	3
Midday Insolation Incident	MIDDAY INSOL	MEAN	MJ/m^2/day	The total amount of solar irradiance (i.e. direct plus diffuse) incident on a horizontal plane at the earth's surface during the solar noon hour midday period.	1
Precipitation Corrected	PRECTOTCORR	MEAN, SUM	mm/day	The bias corrected average of total precipitation at the surface of the earth in water mass (includes water content in snow).	3
Precipitable Water	PW	SUM	cm	The total atmospheric water vapor contained in a vertical column of the atmosphere.	3
Specific Humidity at 10 Meters	QV10M	MEAN	g/kg	The ratio of the mass of water vapor to the total mass of air at 10 meters (g water/kg total air).	3
Specific Humidity at 2 Meters	QV2M	MEAN	g/kg	The ratio of the mass of water vapor to the total mass of air at 2 meters (g water/kg total air).	3

TABLE X: (continued)

Definition	Variable Name	Treatment	Unit	Description	Group
Relative Humidity at 2 Meters	RH2M	MEAN	%	The ratio of actual partial pressure of water vapor to the partial pressure at saturation, expressed in percent.	3
Temperature at 10 Meters	T10M	MEAN	°C	The air (dry bulb) temperature at 10 meters above the surface of the earth.	4
Temperature at 10 Meters Maximum	T10M MAX	MEAN, MAX	°C	The maximum hourly air (dry bulb) temperature at 10 meters above the surface of the earth in the period of interest.	4
Temperature at 10 Meters Minimum	T10M MIN	MEAN, MIN	°C	The minimum hourly air (dry bulb) temperature at 10 meters above the surface of the earth in the period of interest.	4
Temperature at 10 Meters Range	T10M RANGE	MEAN	°C	The minimum and maximum hourly air (dry bulb) temperature range at 10 meters above the surface of the earth in the period of interest.	4
Temperature at 2 Meters	T2M	MEAN	°C	The average air (dry bulb) temperature at 2 meters above the surface of the earth.	4
Dew/Frost Point at 2 Meters	T2MDEW	MEAN	°C	The dew/frost point temperature at 2 meters above the surface of the earth.	3
Wet Bulb Temperature at 2 Meters	T2MWET	MEAN	°C	The adiabatic saturation temperature which can be measured by a thermometer covered in a water-soaked cloth over which air is passed.	4
Temperature at 2 Meters Maximum	T2M MAX	MEAN, MAX	°C	The maximum hourly air (dry bulb) temperature at 2 meters above the surface of the earth in the period of interest.	4
Temperature at 2 Meters Minimum	T2M MIN	MEAN, MIN	°C	The minimum hourly air (dry bulb) temperature at 2 meters above the surface of the earth in the period of interest.	4
Temperature at 2 Meters Range	T2M RANGE	MEAN	°C	The minimum and maximum hourly air (dry bulb) temperature range at 2 meters above the surface of the earth in the period of interest.	4
Air Temperature Range at 2 Meters Maximum	T2M RANGE MAX	MEAN, MAX	°C	The maximum air (dry bulb) temperature range at 2 meters above the surface of the earth in the period of interest.	4
Air Temperature Range at 2 Meters Minimum	T2M RANGE MIN	MEAN, MIN	°C	The minimum air (dry bulb) temperature range at 2 meters above the surface of the earth in the period of interest.	4

TABLE X: (continued)

Definition	Variable Name	Treatment	Unit	Description	Group
Maximum Wet Bulb Globe Temperature	T2M MAX WBG	MEAN, MAX	°C	The maximum Wet Bulb Globe Temperature. The Wet Bulb Globe Temperature (WBGT) is a composite temperature used to estimate the effect of temperature, humidity, wind speed (wind chill), and visible and infrared radiation (usually sunlight) on humans.	4
Minimum Wet Bulb Globe Temperature	T2M MIN WBG	MEAN, MIN	°C	The minimum Wet Bulb Globe Temperature.	4
Temperature Range Maximum	T2M RANGE MAX	MEAN, MAX	°C	The maximum hourly air temperature range (maximum temperature minus minimum temperature).	4
Temperature Range Minimum	T2M RANGE MIN	MEAN, MIN	°C	The minimum hourly air temperature range (minimum temperature minus maximum temperature).	4
Wind Speed at 10 Meters	WS10M	MEAN	m/s	The average of wind speed at 10 meters above the surface of the earth.	5
Wind Speed at 10 Meters Maximum	WS10M MAX	MAX	m/s	The maximum hourly wind speed at 10 meters above the surface of the earth.	5
Wind Speed at 10 Meters Minimum	WS10M MIN	MIN	m/s	The minimum hourly wind speed at 10 meters above the surface of the earth.	5
Wind Speed at 2 Meters	WS2M	MEAN	m/s	The average of wind speed at 2 meters above the surface of the earth.	5
Wind Speed at 2 Meters Maximum	WS2M MAX	MAX	m/s	The maximum hourly wind speed at 2 meters above the surface of the earth.	5
Wind Speed at 2 Meters Minimum	WS2M MIN	MIN	m/s	The minimum hourly wind speed at 2 meters above the surface of the earth.	5
Wind Speed at 50 Meters	WS50M	MEAN	m/s	The average of wind speed at 50 meters above the surface of the earth.	5
Wind Speed at 50 Meters Maximum	WS50M MAX	MAX	m/s	The maximum hourly wind speed at 50 meters above the surface of the earth.	5
Wind Speed at 50 Meters Minimum	WS50M MIN	MIN	m/s	The minimum hourly wind speed at 50 meters above the surface of the earth.	5

Note: Use '_' for spaces in variable names.
The groups are defined as follows:

1 - Solar Radiation: Variables related to solar radiation levels.
2 - Cloud Cover: Variables describing the extent of cloud cover.
3 - Hydrological Variables: Variables associated with water-related factors, such as precipitation and humidity.
4 - Temperature: Variables related to temperature measurements.
5 - Windspeed: Variables describing windspeed and related factors.