Grenze International Journal of Engineering and Technology, January Issue

Predicting Dengue Fever via Weather Data: A Comparative Analysis of Statistical and Machine Learning Models

Vinod Babu P[1], Sri Krishna A[2], RajaSekhar N[3] and Reshma Nawaz MD[4]
[1-4] Department of Artificial Intelligence, Shri Vishnu Engineering College for Women, Bhimavaram, AP.
Email: vinodbabup.ai@svecw.edu.in1, srikrishna.ai@svecw.edu.in2, nrajasekharme@svecw.edu.in3, nawazreshmanz@gmail.com4

***Abstract*— This study, based on meteorological factors, provides a forecast analysis of dengue cases in the Philippines. We used two datasets: one with the number of dengue cases and another with information on the weather such as surface pressure, mean tem- perature, relative humidity, wind speed, precipitation.Multiple machine learning models were utilized to predict future dengue cases: Random Forest classifier, CNN-1D, CNN with an attention mechanism, and Support Vector Regression (SVR). On the DS2 dataset, the CNN with an attention mechanism outperformed the other models, with a Root Mean Square Error (RMSE) of 0.02. Mean temperature is the most important meteorological param- eter impacting the rise in dengue incidence, according to feature importance analysis using SVR. The usefulness of sophisticated deep learning approaches in modeling epidemiological data is shown by this work, which also emphasizes the important role of weather parameters in dengue prediction.This paper highlights the significance of weather factors for dengue prediction and demonstrates the utility of advanced deep-learning algorithms in modelling epidemiological data.**

***Index Terms*— Dengue forecasting, Climatic parameters, CNN with attention mechanism, Epidemiological Modeling.**

I. Introduction

A. Back Ground:

A serious threat to public health in tropical and subtropical regions of the world is dengue fever, which is caused by a virus spread by mosquitoes. Dengue is thought to infect 390 million people annually, making it a major source of illness and death in these regions. The intricate dynamics of dengue transmission are influenced by a multitude of factors, including human behavior, environmental conditions, and the presence of mosquito vectors capable of carrying the virus. The effect of weather on dengue outbreaks is one of these concerns that needs more research. Proactive approaches integrating knowl- edge of the intricate relationship between weather patterns and dengue epidemics are required to prevent and treat illnesses.

B. Why Predictive Analysis Is Important?

Accurate and early forecasting of dengue epidemics has become crucial for public health organizations to implement preventive strategies that work.

Grenze ID: 01.GIJET.11.1.304

Although traditional monitoring methods are important, they often do not provide the timely alarms needed for preventive actions. Adding weather-related factors to forecasting algorithms is one technique to potentially lessen this disparity. Developments in machine learning and data analytics could be used to create models that can quickly adapt to changes in meteorological conditions and simultane- ously identify historical trends. These forecasting algorithms have the power to drastically alter our dengue prevention plan because they enable more precise and thoughtful resource allocation.

C. The Impact of Weather Conditions:

The principal vectors of dengue, Aedes mosquitoes, are highly dependent on the weather throughout their life cycle.In addition to mosquito survival and reproduction, temperature, humidity, wind speed, and precipitation all have an impact on the rate at which viruses propagate among the vectors. As global weather patterns continue to be influenced by climate change, it is projected that the dynamics of dengue transmission will shift, underscoring the need for flexible forecasting models. It is essential to examine the complex connections between weather and dengue cases in order to anticipate future problems related to climate variability and to comprehend existing trends.

II. Related Works

- Historical Perspectives on Dengue Prediction: Most early attempts to forecast Dengue outbreaks relied on statis- tical models, which frequently struggled to capture the complex dynamics of the disease. These models often contained prior case data along with basic epidemiologi- cal parameters. Although perceptive, they were unable to keep up with the dynamic nature of Dengue transmission because they lacked the accuracy necessary for preventive public health measures [1].
- Dengue Machine Learning Applications Machine Learn- ing Utilized in Dengue Machine learning (ML) approaches have been substantially responsible for recent breakthroughs in the prediction of infectious illnesses. Numerous machine learning (ML) models, including as decision trees, neural networks, and support vector ma- chines (SVM), have been investigated in studies to predict dengue epidemics. By adding more variables including socioeconomic indicators, land use patterns, and climatic data, the models' accuracy increased. How these models will adjust to shifting environmental conditions is yet unclear, but [2].
- Spatial and Temporal Aspects in Dengue Prediction: Several studies have combined machine learning models with geographic information systems (GIS) due to the importance of spatiotemporal components in Dengue transmission. These efforts aim to give a more compre- hensive view of the interactions between environmental and demographic factors in specific places by increasing the geographical resolution of Dengue forecasting. De- spite being hopeful, these studies frequently encounter problems with the scalability of spatial modelling tech- niques and data accessibility [3].
- Restrictions and Prospective Paths: The incapacity of the existing ML-based dengue forecasts to incorporate real-time meteorological data limits their reliability to changes in the meteorological conditions. Public health practitioners find it difficult to apply machine learning be- cause of its subtle techniques, which limit interpretability. Future research should enhance technological integration for more precise dengue preventive forecasts. [4].
- Integration of Deep Learning Techniques (2019): New advances in deep learning, including as LSTM networks and RNNs, have the potential to enhance dengue forecast- ing models. Their ability to identify temporal correlations in complex datasets is a valuable asset in the dynamic field of dengue transmission. The challenges of needing enormous volumes of training data and tuning hyperpa- rameters are uncharted territory [5].
- Ensemble Learning Methods (2020): 2020 will be a focus year for research on ensemble learning techniques for dengue prediction. In order to improve overall predic- tion accuracy, a variety of forecasting models, including SVM, decision trees, and neural networks, were incor- porated into an ensemble framework.

  This method aims to increase dengue projection accuracy by reducing the shortcomings of various models [6].

- ➢ Dengue Forecasting with Remote Sensing (2021): Re- search in 2021 concentrated on updating dengue epidemic forecasting models with data from remote sensing. By gathering more environmental data using satellite images and other remote sensing technologies, this strategy seeks to deepen our understanding of the ecological elements impacting Dengue transmission. Nonetheless, there are still issues with data integration and preparation in this developing sector [7].
- ➢ An AI that Explains Dengue Prediction: In 2022, an attempt was made to find a solution to the interpretability
- ➢ problem with sophisticated machine learning models. In an effort to make dengue forecasting models more transparent and intelligible for public health experts, researchers investigated integrating explainable AI approaches into the models. This effort helps close the gap that exists between sophisticated modeling capabilities and real-world implementation [8].

III. DATASET

Two different datasets from the Philippines were painstakingly gathered to create our dengue disease prediction dataset. These datasets are from the same region and are essential to the development of our forecasting model. One kind of dataset that offers details on significant parameters including temperature, precipitation, relative humidity, wind speed, and surface pressure is meteorological data. Simultaneously, the second dataset logs cases of dengue together with the observation date and total number of cases (Cnt) of the illness. Kaggle provided epidemiological data spanning the years 2008 to 2016. Concurrently, meteorological data over the same duration was acquired from the NASA power data website [9]. Variables in this large dataset include temperature, relative humidity, precipitation, wind speed, and surface pressure. A updated version was created following the merger of the two datasets as part of the preparation procedure for the dataset. This second dataset includes delayed time lags to represent the temporal aspects of features. The optimal temporal lags were discovered via cross-validation, resulting in lag values of 1 for surface pressure, 9 for mean temperature, 11 for precipitation, 2 for relative humidity, and 11 for wind speed. Building this second dataset is consistent with the overarching objective of improving dengue forecast accuracy by incorporating temporal factors into the modeling methodology.

	Date	Dengue Count	Mean Temparature	Relative Humidity	Wind Speed	Surface Pressure	Precipitation
0	01-01-2008	79	26.44	83.88	8.27	10.55	101.07
1	01-02-2008	49	26.21	85.00	8.77	5.27	101.06
2	01-03-2008	115	26.64	83.56	5.84	0.00	101.00
3	01-04-2008	131	27.94	82.12	5.70	5.27	100.91
4	01-05-2008	129	27.90	82.12	3.65	10.55	100.69

Fig. 1. Dataset

This combination carefully integrated datasets to match feature significance in cutting-edge machine learning algorithms in an effort to maximize the interplay between Dengue transmission dynamics and weather patterns. The structured dataset, which was centered on the Philippines, aimed to offer a thorough grasp of the relationship between weather patterns and disease occurrence. Its main objective was to improve the accuracy of Dengue forecasts by taking into account the temporal evolu- tion of disease incidence in conjunction with environmental circumstances.

IV. PROPOSED SYSTEM

The proposed methodology for dengue case prediction based on meteorological parameters includes some important steps. Since the data is sequential in nature, it should be stationary. Stationarity was assessed using ADF testing. If the data is stationary, we can proceed to other steps; if not, we can use differencing and box-cox transformation to make the data stationary. The next two datasets to be generated were two. The second dataset, which also includes the number of dengue cases and weather features, has a delayed time lag from the first dataset. For the second batch of data, autocorrelation and cross-correlation are employed. The range of variables or characteristics can be standardized using standard scaling, a preprocessing technique, so that they fall into the

same scale. This scale typically consists of values between 0 and 1, or around a mean of 0 with a standard deviation of 1.
To determine the one factor that is most crucial in increasing the number of dengue cases. We have used the SVR model to do a comprehensive feature significance analysis in addition to prediction. This method adds to a more nuanced under- standing of the many relationships involved by highlighting the significance of each factor in determining dengue outbreaks. Time Series Split () was utilized to divide the data into test and train sets in order to train deep learning models, such as CNN-1D and CNN with attention mechanisms, and time series models, such as SARIMA and ensemble models like Random Forest Classifier. For the purpose of analyzing sequential data, the attention model uses Keras to construct a neural network architecture that combines convolution and recurrent layers with an attention mechanism. With an RMSE score of 0.028, the CNN with an attention mechanism has outperformed the other models overall. CNNs with attention mechanisms offer a robust framework for processing complex data with sequential or spatial structure, hence improving performance. Neural networks can process and focus on the most relevant information thanks to this structure.

V. Methodology

A. Sarima

SARIMA models must be used for the analysis of time series data that exhibit seasonality or recurrent patterns. Com- pared to traditional ARIMA models, they provide a more thor- ough method of forecasting by accounting for both seasonal and non-seasonal elements. However, it is crucial to confirm that the data is stationary before developing an autoregressive model. A series needs to have a constant mean, variance, and covariance in order to be deemed stationar. This can be con- firmed using formal tests that determine if a series is stationary or not, such as the KPSS and ADF Tests. Null Hypothesis (H0)¿ 0.05 occurs when the series in the ADF test is stationary. Differentiating and applying the box-cox transformation are two techniques used to convert non-stationary data into a stationary series. In the SARIMA model, the link between the current and lag observations is reflected by the autoregressive (AR) terms, whereas the moving average (MA) terms capture short-term dependencies. To stabilize the series mean, the integrated (I) terms differ. In addition to seasonal variations, seasonal autoregressive (SAR) and seasonal moving average (SMA) factors are also incorporated in SARIMA to eliminate seasonal trends from the data. Using statistical methods like grid search and optimization algorithms, the SARIMA model parameters (p, d, q) for the seasonal portion and (P, D, Q) for the non-seasonal part are determined. After the model is trained on historical data, it can be used to predict future time points

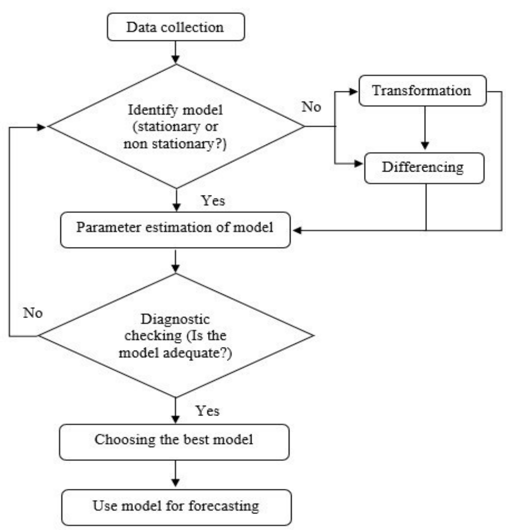

Fig. 2. Architecture of SARIMA Model [11]

B. SVR

The machine learning algorithm Support Vector Regression (SVR) uses the ideas of Support Vector Machines (SVM) for regression analysis. SVR is an excellent method for time series forecasting because it projects the input data into a high- dimensional space and finds a hyper plane that best represents the correlations between the input features and related output values. Unlike conventional regression models, support vector regression (SVR) aims to reduce error inside a predetermined margin or tube around the predicted values. In time series forecasting, SVR uses historical data to find patterns and correlations in order to provide accurate forecasts for future time points.

SVR is particularly helpful for managing intricate and nonlinear interactions in time series data used additionally to identify the meteorological elements that lead to the increase in dengue incidence. By transforming the input into a higher- dimensional space, SVR can capture complicated temporal patterns, which makes it a valuable tool for assignment pre- diction in a variety of industries, such as banking and weather forecasting.

C. Random Forest Classifier

The impact of various meteorological conditions on dengue case prediction was evaluated using the Random Forest classifier, a well-liked ensemble learning technique known for its robustness and interpretability. During training, the Random Forest algorithm generates multiple decision trees, each using a different random subset of the characteristics and data; the outputs of these different trees are combined to produce the final forecast, increasing the model's generalizability and reducing the risk of over fitting. In this paper, the Random Forest classifier was used to forecast the frequency of dengue cases as well as rank the significance of each weather parameter, including mean temperature, relative humidity, wind speed, precipitation, and surface pressure.

By analyzing each feature's contribution, key environmental factors that have a major impact on dengue epidemics were found. The maximum depth and number of trees in the Random Forest model were two major parameters that were adjusted using cross-validation to optimize performance. This method not only helped us comprehend the basic relationships between weather patterns and dengue incidence, but it also provided a helpful baseline against more complex models.

D. CNN-1D

The 1D Convolutional Neural Network (CNN-1D) architecture uses one-dimensional convolutional layers to enhance the detection of connections and trends in sequentially presented data, such time series. CNNs were originally created for image processing, but one-dimensional convolutions have been used to adapt CNNs for time series data.

Procedure in stages

- Typically, the input layer of a 1D CNN receives an array of data points as input. A time interval or feature can be represented by a single data point in a time series dataset.
- Convolutional layers use convolutional filters to process the input data. These filters extract local patterns or features by slicing through the incoming data. The filters in a 1D CNN can only move in one direction along the input sequence. A feature map that highlights specific features or patterns in the data is produced by each filter.
- After the convolution operation, the model is given non- linearity, which allows it to learn complex patterns using an activation function like ReLU (Rectified Linear Unit) element-wise.
- Through the pooling layer, the feature maps generated by the convolutional layers are reduced in dimension. Pooling layers are used to reduce the spatial dimensions of the feature maps while preserving important information. They often use methods like max pooling or average pooling.
- By flattening the output of convolutional and pooling layers into a 1D vector, or reshaping the 2D feature maps, the flattening layer gets the data ready for fully connected layers.
- Fully connected layers capture complex interactions and higher-level abstractions before generating the final predictions. When working with time series data that displays local patterns or when comprehending the spatial links within the sequence is crucial for creating predictions, the output layer creates the final forecast for the time series.

E. CNN with Attention Mechanism

An attention-based CNN architecture is a CNN that combines attention processes with convolutional layers. Attention methods enable the model to focus on specific sequence segments by assigning different weights to different elements of the input sequence based on their significance. This can enhance the network's ability to recognize important patterns and dependencies in time series data.

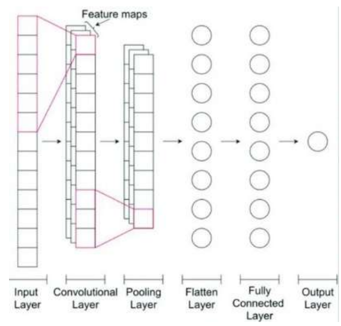

Fig. 3. Architecture of CNN-1D [10]

Procedure in stages

- Time series data, where each data point refers to a feature at a specific interval of time, are fed into the model at the input layer.
- The convolutional layer, which is the following layer, uses filters and/or kernels to extract local patterns or features throughout the temporal dimension of the input data, resulting in the creation of a feature map.
- The Long Short-Term Memory Layer (LSTM Layer) maintains the sequential structure of the time series data and captures longitudinal patterns and trends by maintaining internal states.
- Using LSTM outputs, the Attention Layer determines attention ratings and assigns varying weights to different sequence segments so that the user can concentrate on relevant information. Next, in order to draw attention to the salient features that the LSTM identified, the Attention Combination Layer combines attention scores with LSTM outputs.
- Based on the attention combination layer's output, the flatten layer supplies dense layers with a one-dimensional array as input.
- As flattened data is being used to learn abstract representations, the dense layer activates ReLU to add complexity and non-linearity. Anticipating future values or guessing the value that will appear next in the sequence allows the output layer to produce a continuous value.
- The model's ability to recognize local and temporal patterns in time series data is improved by integrating CNNs, LSTMs, and attention mechanisms. This lets the model perform better when making predictions.

VI. RESULTS AND DISCUSSION

The highest temperature ever recorded, 2 degrees, and the lowest temperature ever recorded, nearly -2 degrees, are shown in graph 3. There is clearly wind because the wind speed ranges from a high of 2 to a low of -2.0. Notable swings may be seen in the relative humidity, which peaked at 90 degrees at the end of 2014 and dropped to lower levels in the first few months of 2015. Dengue cases follow a pattern: after 2014, there was an average rise in instances, with 2015 having the highest number of cases.

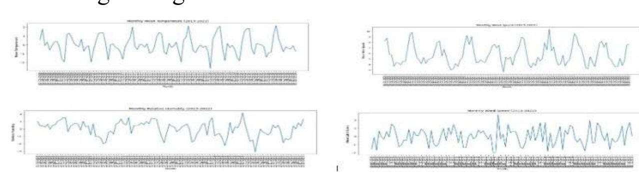

Fig. 4. Time Series plots of climate variables

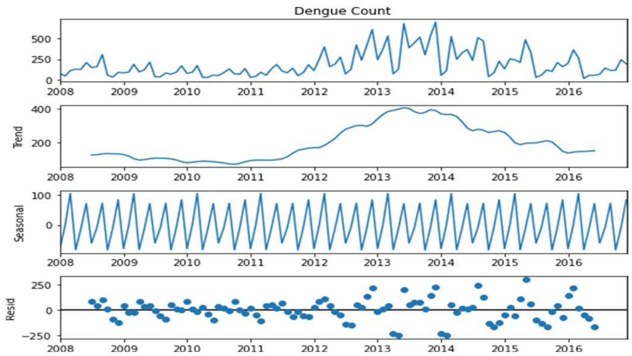

Fig. 5. Time Series Decomposition

The statistics show no discernible trend. The above graph [5] makes seasonality quite evident; while dengue often spreads during the heat and rainy season, we can see that there is a spike in incidence during the middle of the year.

We must run certain statistical tests, such as ADF or KPSS, to determine whether the series is stable or not. Testing has shown that there is a stable relationship between dengue cases and certain meteorological conditions.

```
ADF Test Results for Mean Temparature:
Test Statistic: -4.511611210293464
P-value: 0.00018735554636678452
Critical Values: {'1%': -3.5019123847798657, '5%': -2.892815255482889, '10%': -2.583453861475781}
Is the series stationary? Yes

ADF Test Results for Precipitation:
Test Statistic: -3.2210431902148375
P-value: 0.01879847487523706
Critical Values: {'1%': -3.503514579651927, '5%': -2.893507960466837, '10%': -2.583823615311909}
Is the series stationary? Yes

ADF Test Results for Relative Humidity:
Test Statistic: -5.208299891906708
P-value: 8.455448792193549e-06
Critical Values: {'1%': -3.4942202045135513, '5%': -2.889485291005291, '10%': -2.5816762131519275}
Is the series stationary? Yes

ADF Test Results for Wind Speed:
Test Statistic: -10.838332411041266
P-value: 1.6401678775845948e-19
Critical Values: {'1%': -3.5011373281819504, '5%': -2.8924800524857854, '10%': -2.5832749307479226}
Is the series stationary? Yes

ADF Test Results for Surface Pressure:
Test Statistic: -7.667699418297375
P-value: 1.624542589867005e-11
Critical Values: {'1%': -3.5011373281819504, '5%': -2.8924800524857854, '10%': -2.5832749307479226}
Is the series stationary? Yes

ADF Test Results for Dengue Cases:
Test Statistic: -8.947497173977188
P-value: 8.906474872619528e-15
Critical Values: {'1%': -3.5011373281819504, '5%': -2.8924800524857854, '10%': -2.5832749307479226}
Is the series stationary? Yes
```

Fig. 6. Stationarity Check

Following the use of Box-Cox transformation and differencing, we have discovered that every parameter is stationary[6]. We have utilized SVR for dataset 1 and 2 in order to determine the influencing weather parameter on dengue cases. SVR indicates that Mean Temparature is the factor influencing the dengue cases for lagged dataset and dataset1.

Introducing new methods such as SVR, CNN-1D, CNN with attention mechanism, Random Forest Classifier, and SARIMA.

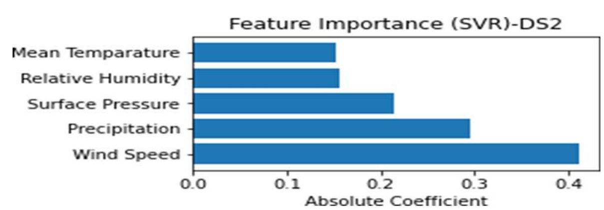

Fig. 7. Feature importance (DS1)

The attention model outperformed all other models in its capacity to predict monthly dengue cases using meteorological factors and dengue cases in the Philippines, with low RMSE scores VI of 0.028 for the lagged dataset (ds2) and 0.07 for ds1.

With regard to time-series forecasting assignments in par- ticular, the noteworthy outcomes of CNN's integration of attention mechanisms highlight the significance of attention processes in enhancing neural network predictive capabilities. By focusing narrowly on relevant temporal characteristics, the model finds and weighs crucial information efficiently and achieves unprecedented precision in refining its predictions. This finding demonstrates the constant evolution of predictive modeling methods as researchers strive to increase prediction accuracy and push performance boundaries.

TABLE I. RMSE VALUES FOR BOTH DS1 AND DS2

Models	**DS1**	**DS2**
SVR	0.975	1.15
SARIMA	0.90	0.92
Random Forest Classifier	1.11	1.20
CNN-1D	1.12	0.77
CNN with attention	0.07	0.028

FUTURE WORK

The sophisticated models CNN with attention, SVR, Grid Search with Gradient Boosting, and SARIMA are integrated into the Dengue illness forecasting architecture, which im- proves accuracy by capturing temporal patterns and complex relationships. There are still more ways to improve it: by retraining, adding different data sources, investigating real- time satellite data, and creating an intuitive user interface specifically for public health professionals. In order to improve prediction and responsiveness to the unpredictability of dengue transmission, cooperation with stakeholders and medical professionals is necessary for a smooth integration into current public health frameworks.

ACKNOWLEDGMENT

The authors acknowledge the support of Department of Science and Technology (DST) for their generous support provided for this research project (Ref. No. DST/SEED/SCSP/STI/2019/140) and Department of Artificial Intelligence, Shri Vishnu Engineering College for Women who gave us the opportunity to use the GPU server at High Performance Computing Lab. We would also like to express our sincere gratefulness to District Health Center, West Godavari for providing the dataset used in this research

REFERENCES

[1] B. Bett, D. Grace, H. S. Lee, J. Lindahl, H. Nguyen-Viet, P.-D. Phuc, N. H. Quyen, T. A. Tu, T. D. Phu, D. Q. Tan et al., "Spatiotemporal analysis of historical records (2001–2012) on dengue fever in vietnam and development of a statistical model for forecasting risk," PLoS One, vol. 14, no. 11, p. e0224353, 2019.

[2] V. J. Jayaraj, R. Avoi, N. Gopalakrishnan, D. B. Raja, and Y. Umasa, "Developing a dengue prediction model based on climate in tawau, malaysia," Acta tropica, vol. 197, p. 105055, 2019.

[3] A. R. A. RM et al., "The design and development of artificial intelligence based 'denv shield'app for predicting dengue: Design of intelligent app 'denv shield'to predict genotype of imminent dengue virus and vector competency of aedes in view of increasing air pollution and climate change," International Journal of Emerging Trends in Science and Technology, vol. 9, 2022.

[4] U. Khan and S. Azeem, "The rising toll of dengue cases in Pakistan every year: An incipient crisis," Annals of Medicine and Surgery, vol. 76, 2022.

[5] C. Davi, A. Pastor, T. Oliveira, F. B. de Lima Neto, U. Braga-Neto, A.

[6] W. Bigham, M. Bamshad, E. T. Marques, and B. Acioli-Santos, "Severe dengue prognosis using human genome data and machine learning," IEEE Transactions on Biomedical Engineering, vol. 66, no. 10, pp. 2861–2868, 2019.

[7] A. Y. Saleh and L. Baiwei, “Dengue prediction using deep learning with long short-term memory,” in 2021 1st International Conference on Emerging Smart Technologies and Applications (eSmarTA). IEEE, 2021, pp. 1–5.
[8] S.-W. Huang, H.-P. Tsai, S.-J. Hung, W.-C. Ko, and J.-R. Wang, “Assessing the risk of dengue severity using demographic information and laboratory test results with machine learning,” PLoS neglected tropical diseases, vol. 14, no. 12, p. e0008960, 2020.
[9] T.-S. Ho, T.-C. Weng, J.-D. Wang, H.-C. Han, H.-C. Cheng, C.-C. Yang, C.-H. Yu, Y.-J. Liu, C. H. Hu, C.-Y. Huang et al., “Comparing machine learning with case-control models to identify confirmed dengue cases,” PLoS neglected tropical diseases, vol. 14, no. 11, p. e0008843, 2020.
[10] https://power.larc.nasa.gov/data-access-viewer/ [10] https://www.mdpi.com/1424-8220/22/23/9517
[11] https://www.researchgate.net/publication/345176326 Comparison of backpropagation artificial neural network and SARIMA in predicting the number of railway passengers