Utilizing Machine Learning Algorithms for Dengue Disease Prediction and Forecasting

Ajay Hiremath
Post grad student at the name
University of Nottingham
MSc in Data Science
Nottingham, UK
ajaychiremathgreprep@gmail.com

Prathamesh Mohan Dharamthok
Post grad student at the name
University of Nottingham
MSc in Data Science
Nottingham, UK
prathamesh.d1998@gmail.com

Pratyush Joshi
Post grad student at the name
University of Nottingham
MSc in Data Science
Nottingham, UK
pratyush.joshi1996@gmail.com

***Abstract*— Dengue fever, an illness transmitted by mosquitoes carrying a virus, presents a significant public health concern, especially in regions with tropical and subtropical climates. The intricate interplay of environmental elements, weather conditions, and human behavior complicates the task of predicting the disease's spread. This study investigates the utility of machine learning methods in forecasting dengue fever outbreaks. Utilizing a dataset containing historical records of climate conditions, population density, and reported dengue cases, we examine various machine learning algorithms, such as linear regression, random forests and support vector machines. Our research aims to evaluate the efficacy of these models in predicting the total number of dengue cases within specific geographical areas over time. Through rigorous experimentation and analysis, we assess the performance of different machine learning techniques, identify crucial factors affecting dengue transmission patterns, and offer insights into effective strategies for disease surveillance and management. The outcomes of this investigation contribute to the ongoing global efforts to combat dengue fever and alleviate its impact on public health.**

Keywords—Dataset, Data preprocessing, Dengue Linear regression, Random Forest, Support vector machine, MAE, RMSE.

I. Introduction and Literature Review

In recent times, the global public health arena has been increasingly influenced by the emergence and spread of infectious diseases. Dengue fever emerges as a particularly formidable challenge, impacting healthcare systems and communities worldwide. As a virus transmitted by mosquitoes, dengue fever crosses geographical borders, affecting tropical and subtropical regions with alarming frequency. Its range of symptoms, from mild flu-like manifestations to severe hemorrhagic fever, underscores the pressing need for effective prevention and control measures.

According to statistics from the World Health Organization (WHO), dengue fever ranks among the fastest-growing mosquito-borne diseases globally, with approximately 390 million infections reported each year [1]. This striking number underscores the pervasive impact of dengue fever on public health, economies, and societal welfare. Moreover, the incidence of dengue fever has seen a significant surge in recent decades, with reported cases increasing twentyfold since the 1960s [2]. This upward trend highlights the escalating threat posed by dengue fever and emphasizes the urgency for concerted efforts to stem its spread. Additionally, the intricate relationship between dengue fever and climatic factors emphasizes the importance of predictive modeling and surveillance. Climate change-induced alterations in temperature, precipitation patterns, and ecological dynamics have profound implications for the distribution and intensity of mosquito-borne diseases such as dengue fever. Thus, precise predictions of disease spread are crucial for informing adaptive public health strategies and mitigating the anticipated consequences of climate-induced epidemiological shifts.

Within this context, our research aims to address the critical need for predictive modeling of dengue fever epidemics. By employing advanced analytical techniques and harnessing comprehensive datasets encompassing epidemiological, climatological, and environmental variables, we aim to unravel the intricate dynamics governing the transmission and propagation of dengue fever. Through our endeavors, we seek to provide policymakers, healthcare professionals, and community stakeholders with actionable insights to proactively combat the threat posed by dengue fever and protect the health and well-being of vulnerable populations.

Following research questions will be addressed as a part of this study:

1. How to deal with data cleaning and filtering of missing values?

2. How do different modeling approaches, such as linear regression, support vector machine and predication using spearman's coefficient impact the performance and interpretability of dengue fever prediction models?

3. How do different machine learning algorithms compare in terms of their predictive performance for dengue fever spread, and which algorithms are most effective in capturing complex relationships in the data?

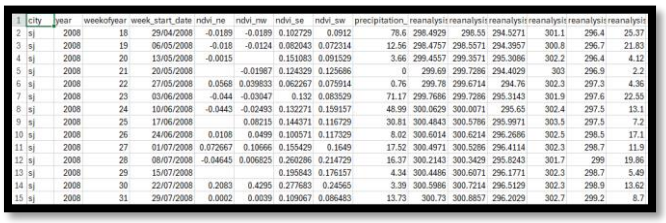
4. What is the importance of features selection for effective prediction?

II. Data

A. Dataset:

The aim is to precisely forecast the total count of dengue fever instances within the test dataset, which will be associated with each city, year, and week of the year. This investigation utilizes information sourced from the DengAI competition (accessible data from the DengAI: Predicting Disease Spread competition on drivendata.org). The DengAI competition provides datasets for two urban areas, San Juan and Iquitos, spanning from three to five years. Each dataset contributes distinct predictive insights for these urban areas. Acknowledging the potential differences in dengue spread patterns between these locations, our strategy entails utilizing

the dataset comprehensively. We will construct separate models for each city and subsequently combine their forecasts to generate our final submission.



	city	year	weekofyear	week_start_date	ndvi_ne	ndvi_nw	ndvi_se	ndvi_sw	precipitation_	reanalysis	reanalysis	reanalysis	reanalysis	reanalysis
2	sj	2008	18	29/04/2008	-0.0189	-0.0189	0.102729	0.0912	78.6	298.4929	298.55	294.5271	301.1	296.4	25.37
3	sj	2008	19	06/05/2008	-0.018	-0.0124	0.082043	0.072314	12.56	298.4757	298.5571	294.3957	300.8	296.7	21.83
4	sj	2008	20	13/05/2008	-0.0015		0.151083	0.091529	3.66	299.4557	299.3571	295.3086	302.2	296.4	4.12
5	sj	2008	21	20/05/2008		-0.01987	0.124329	0.125686	0	299.69	299.7286	294.4029	303	296.9	2.2
6	sj	2008	22	27/05/2008	0.0568	0.039833	0.062267	0.075914	0.76	299.78	299.6714	294.76	302.3	297.3	4.36
7	sj	2008	23	03/06/2008	-0.044	-0.03047	0.132	0.083529	71.17	299.7686	299.7286	295.3143	301.9	297.6	22.55
8	sj	2008	24	10/06/2008	-0.0443	-0.02493	0.132271	0.159157	48.99	300.0629	300.0071	295.65	302.4	297.5	13.1
9	sj	2008	25	17/06/2008		0.08215	0.144371	0.116729	30.81	300.4843	300.5786	295.9971	303.5	297.5	7.2
10	sj	2008	26	24/06/2008	0.0108	0.0499	0.100571	0.117329	8.02	300.6014	300.6214	296.2686	302.5	298.5	17.1
11	sj	2008	27	01/07/2008	0.072667	0.10666	0.155429	0.1649	17.52	300.4971	300.5286	296.4114	302.3	298.7	11.9
12	sj	2008	28	08/07/2008	-0.04645	0.006825	0.260286	0.214729	16.37	300.2143	300.3429	295.8243	301.7	299	19.86
13	sj	2008	29	15/07/2008			0.195843	0.176157	4.34	300.4486	300.6071	296.1771	302.3	298.7	5.49
14	sj	2008	30	22/07/2008	0.2083	0.4295	0.277683	0.24565	3.39	300.5986	300.7214	296.5129	302.3	298.9	13.62
15	sj	2008	31	29/07/2008	0.0002	0.0039	0.109067	0.086483	13.73	300.73	300.8857	296.2029	302.7	299.2	8.7

FIGURE I: AN OVERVIEW OF DATASET

B. Data Manipulation:

In this step, we manipulate the data to make it efficient. Data cleansing involves identifying and rectifying inaccurate, incomplete, or inconsistent data. In this context, inconsistency may arise from misspellings or errors in data originating from a singular source, necessitating the correction of improper data. Data integration entails merging data from diverse sources often presented in different formats, requiring consolidation or integration to eliminate unnecessary or duplicate data [13]. Data scrubbing or cleansing is integral to predictive modeling, as it ensures the reliability of data relationships among variables. Traditional methods often yield suboptimal results when analyzing data relationships, highlighting the importance of implementing a dependable system capable of handling poor-quality data.

B. Data Preprocessing:

1. Visualization of Missing values:
After analyzing the data, we found that certain rows have missing values. For data preprocessing we have implemented removal of rows having more than 3 missing values.
Below graph visualizes the number of missing values in the data set.

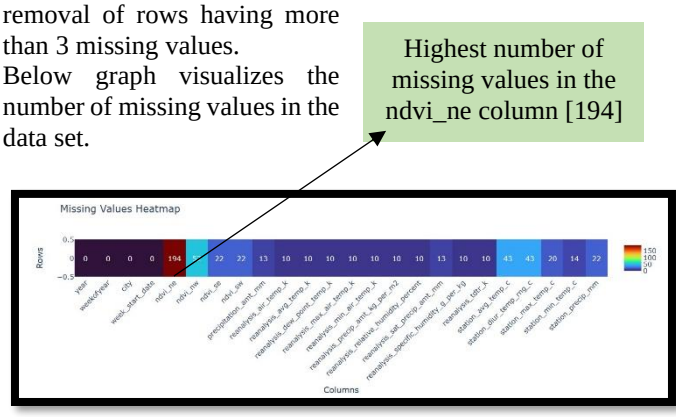


FIGURE II: NUMBER OF MISSING VALUES IN A COLUMN.

2. Handling Missing values:
To deal with the missing values we have used the KNN ("K-Nearest Neighbors") imputation method. KNN imputation is a method for filling in missing data by examining similar instances. It finds the closest neighbors to a data point with missing values, measures their likeness, and computes a weighted average of their known values to predict the missing one.

C. Data splitting:
Data splitting is crucial in model formation, we need to split the data into separate datasets in a rational way.

1. Training Set (80%): This subset comprises 80% of the original dataset and is used to train your machine learning model. The model learns patterns, relationships, and features from this portion of the data.

2. Validation Set (20%): This subset comprises 20% of the original dataset and is held out from the training process. It is used to evaluate the performance of your trained model on unseen data. The testing set provides an unbiased estimate of how well your model generalizes to new, unseen observations.

By splitting the dataset into these two subsets, we can effectively train and evaluate the machine learning model, ensuring that it learns from a diverse range of examples and can make accurate predictions on new data. The 80-20 division ensures a harmonious allocation of data for both training the model effectively and assessing its performance accurately. By doing so, it guards against overfitting, where the model excessively memorizes the training data and fails to generalize to unseen data. This split also facilitates dependable evaluation of the model's performance.

III. METHODOLOGY

To approach this problem, we have implemented different algorithms to get the efficient result. As observed from the above data, there are many features which may or may not have an impact on the result. Different models will yield different results and will give us an insight on the efficiency.

1st Approach: Selecting all the features:

Beginning with considering all the features we try to form different models.

a. Random Forrest Model:
- Random Forest is an ensemble learning technique that amalgamates multiple decision trees to make predictions.
- Random Forests are resilient against overfitting and can capture intricate relationships between predictors and the target variable.
- Training MAE: 4.32, Training RMSE: 8.20
- Validation MAE: 10.853, Validation RMSE: 19.89
- The Random Forest model demonstrates strong performance on the training data, with relatively low mean absolute error (MAE) and root mean squared error (RMSE) values. However, there is a noticeable increase in error metrics when applied to the validation data, indicating some degree of overfitting. Further tuning of model hyperparameters or feature selection may be beneficial to improve generalization performance.

b. Linear Regression Model:
- Linear Regression is a simple yet powerful statistical method for modelling the relationship

between variables. It assumes a linear relationship between predictors and the target variable.

- Training MAE: 15.79, Training RMSE: 26.8
- Validation MAE: 16.402, Validation RMSE: 26.91
- The Linear Regression model exhibits higher error metrics compared to the Random Forest model, suggesting that it may not capture the underlying complexity of the data as effectively. The noticeable gap between training and validation error metrics indicates potential issues with model generalization. Feature engineering or the use of more advanced modelling techniques could be explored to improve performance.

c. Support Vector Machine (SVM):

- SVM is a supervised learning algorithm that finds the optimal hyperplane to separate data points into different classes. It can handle both linear and nonlinear relationships between variables.
- Training MAE: 17.02, Training RMSE: 32.17
- Validation MAE: 13.57, Validation RMSE: 28.47
- The SVM model demonstrates relatively consistent performance between training and validation datasets, indicating good generalization ability. However, the error metrics are higher compared to the Random Forest model, suggesting that the SVM may not capture the complexity of the data as effectively. Further experimentation with different kernel functions or regularization techniques may be warranted to improve performance.

2nd Approach: Feature selection using Spearman's rank coefficient.

Spearman's rank coefficient, also known as Spearman's correlation coefficient or Spearman's rho, is a non-parametric measure of correlation between two variables. Spearman's rank coefficient assesses monotonic relationships, meaning that it captures any consistent increase or decrease in the variable's values, regardless of their linearity.

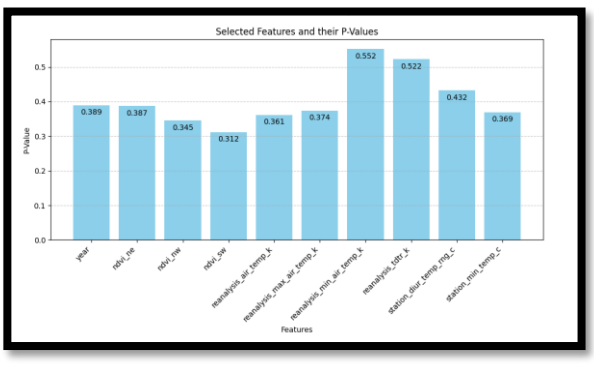


FIGURE III: PLOT OF SELECTED FEATURES USING SPEARMAN'S RANK COEFFICIENT

Spearman's rank coefficient is calculated by first ranking the values of each variable and then computing the Pearson correlation coefficient between the ranked variables. It ranges from -1 to 1, where -1 indicates a perfect negative monotonic relationship, 1 indicates a perfect positive monotonic relationship, and 0 indicates no monotonic relationship.

We employed feature selection using Spearman's rank coefficient to identify the most relevant features for predicting dengue fever cases. The selected features and their corresponding p-values as shown in Figure III:

Comparison of Model Results:

When comparing the results of the models trained with all features versus the models trained with selected features using Spearman's rank coefficient, several observations can be made:

- Random Forest: The Random Forest model exhibits similar performance between the two approaches, with slightly lower MAE and RMSE values on the training and validation datasets when using selected features.
- Linear Regression: The Linear Regression model shows noticeable improvements in MAE and RMSE values on the validation dataset when using selected features, indicating that feature selection helped in reducing prediction errors.
- Support Vector Machine (SVM): The SVM model also demonstrates improved performance on the validation dataset with selected features, particularly in terms of MAE, suggesting that feature selection contributed to better predictive accuracy.

Overall, feature selection using Spearman's rank coefficient appears to have a positive impact on model performance, especially for Linear Regression and SVM models, by focusing on the most relevant features for predicting dengue fever cases.

Model	Training MAE	Training RMSE	Validation MAE	Validation RMSE
Random Forest	4.97	8.97	5.10	8.50
Linear Regression	16.10	27.20	16.54	27.39
Support Vector Machine	16.95	32.11	16.43	32.00

TABLE I: MODEL PERFORMANCE

This table presents the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for each model on both the training and validation datasets. It provides a concise overview of the model performance metrics obtained in the experiment.

3rd Approach: Selecting top 10 features:

The third approach focuses on selecting the top 10 most important features for each machine learning model. This approach recognizes that not all features contribute equally to the predictive performance of a model, and by selecting only the most relevant features, we can potentially improve the model's accuracy and efficiency.

Feature Selection: The first step involves identifying the top 10 important features for each model. This is typically done using feature importance techniques specific to each model. For example, Random Forest models often provide feature importance scores based on how much each feature decreases the impurity of the nodes in the trees, while linear regression models might use coefficients or statistical tests to determine feature importance.

Model Training and Evaluation: Once the top features are selected, the machine learning models are trained and evaluated using only these features. This helps in reducing the dimensionality of the dataset and focusing the model's attention on the most informative features.

Performance Evaluation: Finally, the performance of each model is evaluated using standard metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). These metrics provide insights into how well the models can predict the target variable (in this case, likely the number of dengue fever cases) using the selected features.

Selected Features:

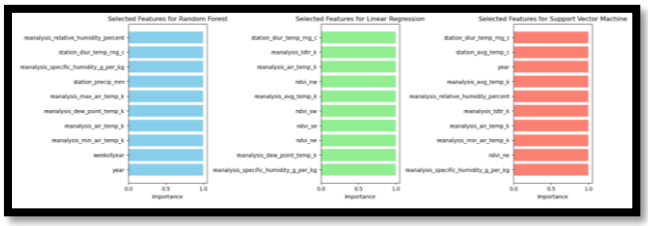


FIGURE IV: PLOT OF TOP 10 SELECTED FEATURES FOR EACH MODEL

The results indicate the performance of each model in terms of Mean Absolute Error and Root Mean Squared Error.

Results for Training dataset:

Mean Absolute Error for Random Forest: 3.98

Mean Absolute Error for Linear Regression: 16.41

Mean Absolute Error for Support Vector Machine: 17.09

Root Mean Squared Error for Random Forest: 7.64

Root Mean Squared Error for Linear Regression: 27.53

Root Mean Squared Error for Support Vector Machine: 32.13

Results for Validation dataset:

Mean Absolute Error for Random Forest: 9.79

Mean Absolute Error for Linear Regression: 16.41

Mean Absolute Error for Support Vector Machine: 17.09

Root Mean Squared Error for Random Forest: 18.70

Root Mean Squared Error for Linear Regression: 27.90

Root Mean Squared Error for Support Vector Machine: 32

IV. RESULTS

The results obtained from the 3 approaches are compared using visualization as shown below:

Results from 1st approach:

The histogram compares the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) values for different models across different approaches. Each bar in the histogram represents a model, and the height of the bars indicates the error values. The histogram provides a visual comparison of the errors for each model, allowing easy identification of the best-performing model for each approach.

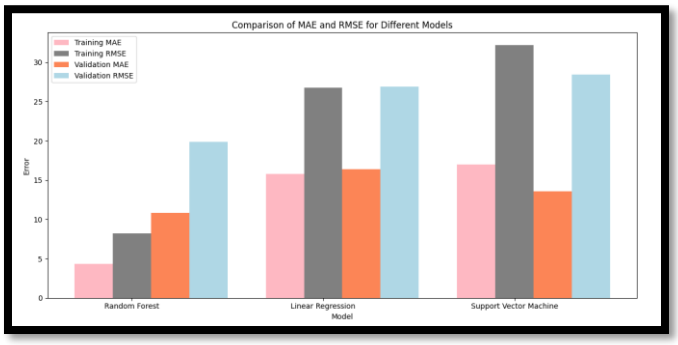


FIGURE V: HISTOGRAM FOR RESULTS FROM 1ST APPROACH

Results from 2nd approach:

The pie chart illustrates the distribution of total validation errors (sum of validation MAE and RMSE) for each model in the second approach. Each slice of the pie represents a model, and the size of the slice corresponds to the proportion of total errors attributed to that model. The pie chart offers a clear overview of the relative contribution of each model to the total validation errors, facilitating comparison and identification of the most impactful models.

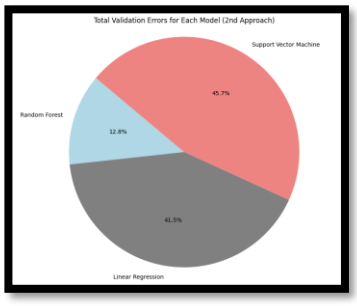


FIGURE VI: PIE CHART FOR RESULTS FROM 2ND APPROACH

Results from 3rd approach:

The line graph displays the training and validation MAE/RMSE values for each model in the third approach of selecting the top 10 features. Separate lines are plotted for training and validation errors, with markers indicating the data points.

The line graph enables comparison of training and validation performance for each model, highlighting any discrepancies and providing insights into the model's generalization capability.

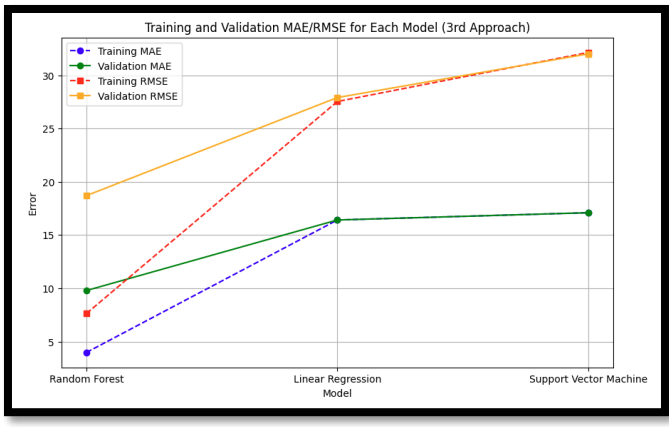


FIGURE VII: LINE GRAPH FOR RESULTS FROM 3RD APPROACH

V. DISCUSSION

1. First Approach: Employing All Features.

Findings Interpretation: This approach involved using all available features for model training and prediction. It aimed to capture relationships between variables and the target (dengue fever cases). The findings focused on assessing predictive performance and identifying patterns from the entire feature set.

Comparing Results: Results were compared with baseline models or prior studies to assess the effectiveness of using all features. Discussions may have covered computational efficiency, model complexity, and the risk of overfitting due to the high-dimensional feature space.

2. Second Approach: Spearman's Rank Coefficient for Feature Selection.

Findings Interpretation: This method selected features based on their correlation with the target variable using Spearman's rank coefficient. Interpretations highlighted the importance of selected features in predicting dengue fever cases and understanding climate-disease relationships.

Comparing Results: Outcomes were compared with the first approach to evaluate the impact of feature selection on model performance. Discussions also addressed the effectiveness of Spearman's rank coefficient in feature selection, along with its limitations.

3. Third Approach: Identifying Top 10 Features.

Findings Interpretation: This approach focused on selecting the ten most significant features for each model using techniques like feature importance scores, coefficient scores. Interpretations emphasized the key predictors of dengue fever cases identified by each model.

Comparing Results: Findings were compared with those from previous approaches to assess the effectiveness of feature selection. Discussions considered model complexity, interpretability, and the generalizability of findings based on selected features.

Each approach provided unique insights into predictive modeling of dengue fever spread, considering feature selection, model performance, and interpretability. Comparing results from different approaches helped identify effective strategies for predicting and understanding disease dynamics.

VI. CONCLUSION

This study explored three different methods for predicting dengue fever spread: utilizing all features, employing Spearman's rank coefficient for feature selection, and identifying the top 10 features for each model.

Using All Features: Employing all available features for model training produced moderate predictive performance, with a Mean Absolute Error (MAE) of approximately 16.39 and a Root Mean Squared Error (RMSE) of around 27.49. While offering a comprehensive understanding of the data, this approach may have led to potential overfitting due to the high-dimensional feature space.

Spearman's Rank Coefficient for Feature Selection: Selecting features based on their correlation with the target variable showed some improvement in model performance, resulting in a lower MAE of approximately 16.12 and RMSE of around 27.19. However, it might have overlooked important features without strong linear relationships with the target variable.

Identifying Top 10 Features: Focusing on the top 10 most significant features for each model enhanced interpretability and possibly reduced overfitting. This approach resulted in a lower MAE of about 9.96 and RMSE of around 19.28. However, the selection process might have been influenced by the modelling techniques used, introducing potential biases.

Consideration of Limitations: Despite promising results, several limitations must be acknowledged. Reliance on predefined feature selection methods may have limited the exploration of novel data relationships. Additionally, predictive performance could have been affected by data quality, completeness, and unaccounted external factors.

Suggestions for Future Research: Future studies could benefit from hybrid approaches that integrate different feature selection techniques to leverage domain knowledge and data-driven insights. Additionally, exploring advanced machine learning algorithms and ensemble techniques may further enhance predictive accuracy and robustness.

In summary, while our study provides valuable insights into dengue fever prediction, there's scope for improvement and exploration of new methodologies. Addressing limitations and embracing innovative approaches can contribute to more accurate and actionable predictions in disease epidemiology.

VII. REFERENCES

[1] A. Wilder-Smith, and P. Rupali, “Estimating the dengue burden in
India,” The Lancet Global Health, vol. 7, pp .e988-e989, 2019.

[2] M. Chovatiya, A. Dhameliya, J. Deokar, J. Gonsalves, and A. Mathur,
“Prediction of Dengue using Recurrent Neural Network,” IEEE International Conference on Trends in Electronics and Informatics (ICOEI), pp. 926-929, 2019.

[3] P. Guo, Q. Zhang, Y. Chen, J. Xiao, J. He, Y. Zhang, L. Wang, T.
Liu, and W. Ma, An ensemble forecast model of dengue in Guangzhou, China using climate and social media surveillance data,” *Science of The Total Environment*, vol. *647*, pp.752-762, 2019.

[4] P. Chen, X. Fu, S. Ma, H.Y. Xu, W. Zhang, G. Xiao, R. Siow Mong
Goh, G. Xu, and L. Ching Ng, “Early dengue outbreak detection modeling based on dengue incidences in Singapore during 2012 to 2017,” Statistics in Medicine, 2020.

[5] O. Şerban, N. Thapen, B. Maginnis, C. Hankin, and V. Foot, “Realtime
processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification”, Information Processing & Management, vol. 56, pp. 1166-1184, 2019.

[6] J.M. Scavuzzo, F. Trucco, M. Espinosa, C.B. Tauro, M. Abril, C.M.
Scavuzzo, and A.C. Frery, “Modeling Dengue vector population using remotely sensed data and machine learning.,” Acta tropica, vol.
185, pp.167-175, 2018.

[7] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease prediction by machine learning over big data from healthcare communities. IEEE Access, vol. 5, pp. 8869-8879, 2019.

[8] J. Andreu-Perez, C. C. Poon, R. D. Merrifield, S. T. Wong, and G. Z.
Yang, “Big data for health,” IEEE journal of biomedical and health informatics, vol. 19, pp. 1193-1208, 2015.

[9] H. Abe, Y. Ushijima, M.M. Loembe, R. Bikangui, G. Nguema-Ondo,
P.I. Mpingabo, V.R. Zadeh, C.M. Pemba, Y. Kurosaki, Y. Igasaki, and S.G. de Vries, “Re-emergence of dengue virus serotype 3 infections in Gabon in 2016–2017, and evidence for the risk of repeated dengue virus infections,” International Journal of Infectious
Diseases, 91, pp.129-136, 2020.

[10] A. Husnayain, A. Fuad, and L. Lazuardi, “Correlation between Google Trends on dengue fever and national surveillance report in Indonesia,” Global Health Action, vol. 12, no. 1, pp.1552652, 2019.

[11] G. J. Milinovich, S. M. Avril, A. C. Clements, J. S. Brownstein, S.
Tong, and W. Hu, Using internet search queries for infectious disease
surveillance: screening diseases for suitability. BMC infectious diseases, vol. 14, pp .690, 2014.

[12] A. Anitha, and D. J. W. Wise, “Forecasting Dengue Fever using Classification Techniques in Data Mining”, International Conference
on Smart Systems and Inventive Technology (ICSSIT) pp. 398-401, 2018.

[13] S. A. Ahmed, and J. S. Siddiqui, “Principal component analysis to
explore climatic variability and dengue outbreak in lahore”, Pakistan Journal of Statistics and Operation Research, vol. 10, pp. 247-256, 2014.

[14] M. Cabrera, and G. Taylor, “Modelling spatio-temporal data of dengue fever using generalized additive mixed models,” Spatial and spatio-temporal epidemiology, vol. 28, pp.1-13, 2019.

[15] M. Mishra, P.B. Dash, J. Nayak, B. Naik, and S.K. Swain, “Deep
Learning and Wavelet Transform integrated approach for Short-term
solar PV power prediction,” Measurement, p.108250, 2020.

[16] B.K. Acharya, C. Cao, T. Lakes, W. Chen, S. Naeem, and S. Pandit,
“Modeling the spatially varying risk factors of dengue fever in Jhapa district, Nepal, using the semi-parametric geographically weighted regression model,” International journal of biometeorology, vol. 62, no. 11, pp.1973-1986, 2018.

[17] S. Demigha, “A case-based reasoning tool for breast cancer knowledge management with data mining concepts and techniques”, In Medical Imaging 2016: PACS and Imaging Informatics: Next Generation and Innovations,” International Society for Optics and Photonics, vol. 9789, pp. 97890I, 2016.

[18] K. Phakhounthong, P. Chaovalit, P. Jittamala, S. D. Blacksell, M. J.
Carter, P. Turner, K. Chheng, S. Sona, V. Kumar, N. P. Day, and L. J.
White, “Predicting the severity of dengue fever in children on admission based on clinical features and laboratory indicators: application of classification tree analysis”, BMC pediatrics, vol. 18, pp. 109, 2018.