Leveraging Machine Learning for Dengue Prevalence Prediction

Aabid Mushtaq Najar[1], Arpan Bhattacharya[2], Shivam Sharma[3] Bansi Dhar Jha[4] Hashim Zahoor

[1] Assistant Professor, School of Computer Applications, Lovely Professional University, Punjab, India
[2]Research Scholar, School of Computer Applications, Lovely Professional University, Punjab, India
[3] Research Scholar, School of Computer Applications, Lovely Professional University, Punjab, India
[4] Research Scholar, School of Computer Applications, Lovely Professional University, Punjab, India
[5] Assistant Professor, School of Computer Applications, Lovely Professional University, Punjab, India

[1]aabidrahie@gmail.com*; [2]bhattacharyaarpan50@outlook.com; [3]shivamsharma1822002@gmail.com, [4]bansi805419@gmail.com; [5]hashimzahoor46@gmail.com

Abstract: Dengue fever is a serious health issue worldwide that is particularly common in warm climates. The diagnosis is based on the identification of symptoms such as fever, headache, joint pain, and rash, which are verified by laboratory testing. Dengue cases are frequently underreported, according to studies that emphasize the need for improved tracking. This study examined dengue hazards and prevalence in the rural and urban areas of India. Hospitals, surveys, and laboratories provided the data. High dengue rates, particularly among young people in metropolitan areas, are associated with characteristics such as high population density and inadequate sanitation. Most illnesses occur during the rainy season. To reduce dengue, focused initiatives such as improved waste management and education are essential. This study will help to improve strategies for preventing and managing dengue fever.

Keywords— **Google Form, Data Pre-processing, Machine Learning, Hospital**

I. INTRODUCTION

Dengue! In tropical regions, the dengue virus causes this fever. The largest issue in the world is fever, which is primarily observed in tropical and warm regions. It is a virus/severe illness brought on by mosquito bites (particularly by the Aedes aegypti species) [1]. People may not become ill immediately after being bitten; it may take a day or two for the virus to take effect. After two days, people start to feel ill with symptoms such as high fever, headaches, and joint and muscle pain, as well as sudden red skin rashes. This fever can occasionally be fatal. Because dengue symptoms are similar to those of other illnesses such as the flu, doctors find it difficult to make a diagnosis. It is difficult to determine the precise number of persons seeking the correct diagnosis, even though doctors typically examine these symptoms and perform certain tests. Indian medical scientists are conducting research to learn more about this fever [2-3]. They are gathering patient data from multiple hospitals, speaking with villagers or those who have been afflicted by dengue, and finally examining lab results to determine the number of dengue cases reported and the region most affected. The purpose of this study is to determine what causes some people to be impacted more frequently than others and what kind of treatment would be best for the patient at that particular moment. By solving this, we seek to keep society healthier and happier, while also identifying the best and most effective strategies to prevent the spread of dengue [4-5].

II. RELATED WORK

To determine the extent to which dengue has spread in particular regions, we created a survey using Google Forms. This survey form was distributed to residents of the study areas. The primary goal of the survey was to identify potential influencing factors and inform respondents of the number of people afflicted with the potentially fatal dengue virus. These variables may include factors such as the age of the individuals or children, their residence, recent vacation destinations, and their surroundings. Respondents provided information about themselves, their residence, and any recent hardships experienced by friends, relatives, or neighbors. Based on current studies on dengue and its transmission from person to person, we created survey questions. We asked respondents about their age, gender, number of people living in their homes, level of cleanliness they maintain, how tidy and clean they

M. Sharma et al. (eds.), *Proceedings of the Third International Conference on Recent Advances in Computing Sciences (RACS 2025)*, Advances in Intelligent Systems Research 212,
https://doi.org/10.2991/978-94-6239-768-2_24

keep their belongings, and other pertinent questions about dengue fever. We also attempted to determine whether anyone had previously experienced any dengue symptoms or had ever had this illness. Experts in public health and disease outbreaks reviewed the questions to ensure that the survey provided accurate data. We also evaluated the survey using a sample of respondents to check whether any questions were unclear. We circulated this survey to local health clinics, community associations, and social media platforms, such as Facebook, Instagram, and Twitter, after everything was prepared. Following the completion of the survey, we gathered all the replies and created a spreadsheet in which we could analyze the information and ensure that it was correct, consistent, and clean. We also ensured that all the responses were formatted in the same way, which involved several steps, including, Data cleaning is the process of eliminating duplicate entries, incomplete responses, and inaccurate or irrelevant data.

1. Data transformation is the process of transforming unprocessed data into a format that can be easily understood.
2. Data integration is the process of merging information from several sources to produce a distinct dataset for additional analysis.
3. Outlier detection refers to data that is either far from the actual data or insufficient from the data that is provided. Outliers can be dealt with by applying statistical techniques to stop them from affecting the analysis's findings.
4. Data validation is the process of confirming the accuracy of data by cross-referencing it with other sources.
5. Data normalization means normalization of variables to ensure comparability and facilitate meaningful comparisons across different data subsets or study populations.
6. Sensitivity Analysis refers to carrying out sensitivity assessments to evaluate how resilient the findings are to changes in data pre-processing and analytical methodologies.
7. Quality Control Checks refer to the use of quality control checks to find and fix outliers, inconsistencies, and data input mistakes that can distort the analysis's findings.
8. Addressing Bias means examining potential sources of bias, such as selection or response bias, and implementing measures to reduce their impact on the study populations.

Moreover! We used more than just surveys to cross-check the results; we also randomly picked a few survey respondents and conducted interviews with them to learn more about their issues. We learned more information during the in-depth discussion on this issue that was not obtainable through the survey forms, which enabled us to comprehend the social and cultural elements as well as the behaviors of individuals that impact the transmission of dengue fever from one person to another.

Our study obtained a comprehensive picture of the prevalence of dengue fever in this region by reviewing previous studies, purifying survey data, and integrating the findings from in-depth interviews. We researched the diverse elements that have a greater impact on people, and a comprehensive approach not only increases the reliability of our findings but also demonstrates our ability to manage teams in a variety of domains. By making decisions about complex public health challenges like dengue based on a variety of evidence.

III. Proposed Methodology Architecture Datasets

Dengue Epidemiological Data: This type of data includes variables such as demographics, environmental factors, dengue incidence rates, and potential risk factors for dengue transmission. It is gathered from various sources, including healthcare, community surveys, and laboratory reports.

3.1. Data Pre-processing:

Cleaning includes eliminating duplicate entries, correcting errors or inconsistencies in the data, and managing missing values using imputation or deletion approaches. Transformation: Translates unprocessed data into comprehensible data. Integration of data from many sources to produce a distinct dataset for analysis is known as integration. The term \"outlier detection\" refers to data that is either insufficient or far from the actual data. Validation is the process of confirming the accuracy of data by cross-referencing it with other sources or comparing it to predetermined standards [7-8].

3.1.1. Feature Selection:

The process of identifying which individual features are most relevant to the target variable by assessing their significance using statistical tests such as ANOVA or the chi-square test is known as univariate feature selection. The Importance of Features Techniques methods for applying tree-based models or feature importance algorithms like these features are ranked based on how much they enhance model performance using Random Forest Feature Importance. Using methods such as Principal Component Analysis (PCA) or feature embedding, dimensionality reduction is the process of lowering the dimensionality of a dataset while preserving relevant information.
The quantity of horizontal edge transitions in pixel intensity values along the horizontal lines within the signature is referred to as the horizontal edges. The signature's horizontal structure or texture can be measured by counting the number of horizontal edges, and vertical edges denote transitions in pixel intensity values along the vertical lines within the signature. Counting the number of vertical edges provides insight into the vertical structure or texture of the signature.

1. Number of Edge Points: Edge points are the pixels where significant changes in intensity occur, indicating boundaries or transitions within the signature. Counting the number of edge points helps assess the overall sharpness or clarity of the signature.
2. Number of Lines (Horizontal and Vertical): Lines represent prominent straight segments within the signature, either horizontally or vertically oriented. Counting the number of horizontal and vertical lines helps characterize the overall structure and organization of the signature.
3. Number of Branch Points: Branch points are places where several edges diverge or connect inside the signature. The complexity and branching structure of the signature can be inferred by counting the number of branch points.
4. These features collectively capture various aspects of the signature's size, shape, texture, and complexity, facilitating its analysis and classification in signature verification systems.
5. The following actions are conducted to extract the geometric feature vector from a 256 × 512 pre-processed signature image.
6. Global Feature Extraction: The global features are ten features that are taken from the complete signature image.
7. Local Feature Extraction: To acquire local features, the signature image is divided into four equal pieces, and the same ten features are taken out of each segment.

Consequently, the signature image yields a geometric feature vector of length 50 after 10 global and 40 local features are removed [9-10].

3.1.2. Splitting the Data:

The pre-processed dataset is split into two parts: a training set and a testing set. The amount of data typically used is 80% for training the machine learning models, and 20% is used for testing and evaluating the performance of the machine learning models. This segmentation ensures good generalization of the developed model to unseen data and reduces the risk of overfitting.

3.2. Machine Learning Models:

a. Random Forest is a tree-based learning method that handles nonlinear relationships and improves prediction accuracy by combining many decision trees.
b. Regression and classification problems are addressed using a supervised learning algorithm technique known as K-Nearest Neighbors (KNN). When working with large datasets, the incredibly powerful XGBoost algorithm, also known as the Extreme Gradient Boosting algorithm, is used to optimize the efficacy and efficiency of gradient boosting techniques.
c. Support Vector Machine (SVM): This method can be used for linear or non-linear classifications, regression, and even outlier detection applications.

3.3. Model Training and Evaluation

Techniques such as grid search or random search are used to train each ML model on the training dengue dataset to optimize the model performance. Once the ML model has been fully trained, its performance is evaluated using a separate dataset to measure a number of skills such as area under the receiver operating characteristic curve (AUC-ROC), accuracy, precision, recall, and F1-score [11-12].

Note: Using this approach, we intend to develop robust and powerful machine-learning models that can accurately predict dengue outbreaks and prevalence. These models should also be able to identify areas with the highest risk factors for virus transmission. This approach integrates data preparation, feature selection, model selection, and evaluation processes to ensure the precision and effectiveness of the prediction models in addressing the challenges posed by dengue.

a) Training-Validation-Testing Split: To effectively assess the model's performance, the dataset was divided into training, validation, and testing sets.

b) Stratified Sampling: strategies to ensure that each subset maintains the same class distribution as the original dataset; this is particularly important for unbalanced datasets.

c) Stratified Sampling: strategies to ensure that each subset maintains the same class distribution as the original dataset; this is particularly important for unbalanced datasets.

IV. RESULTS AND DISCUSSION

The competition between several machine learning models and an examination of each model's expected accuracy using the previously mentioned metrics were the results of the proposed methodology for estimating dengue prevalence. The XGBoost algorithm predicted dengue virus outbreaks with a high accuracy of 91.5% using the survey data provided. Each ML model has the ability to anticipate virus outbreaks, as evidenced by the potential variations in the models' performance, such as their overall prediction accuracy.

- The performance of each machine learning model (Random Forest, KNN, XG-Boost, and SVM) was evaluated using measures such as the F1-score, accuracy, precision, recall, and AUC-ROC.
- Despite variations in performance metrics, all models demonstrated relatively high levels of accuracy, indicating their potential utility in dengue prediction tasks.
- The XG-Boost model outperformed other models with an accuracy of 91.5% on the testing dataset, demonstrating its effectiveness in forecasting dengue prevalence based on the selected characteristics.

The performance of the various machine learning models was assessed using a number of metrics, including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC), as described above. The results are summarized in Table 1. The superiority of the XG-Boost model was demonstrated by its high accuracy of 91.5%.

Table 1: Comparisons of Multiple Models for Performance Parameters

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC-ROC**
Random Forest	89.2%	0.87	0.91	0.89	0.92
KNN	86.5%	0.85	0.87	0.86	0.89
XG-Boost	91.5%	0.92	0.90	0.91	0.94
SVM	88.7%	0.88	0.86	0.87	0.9

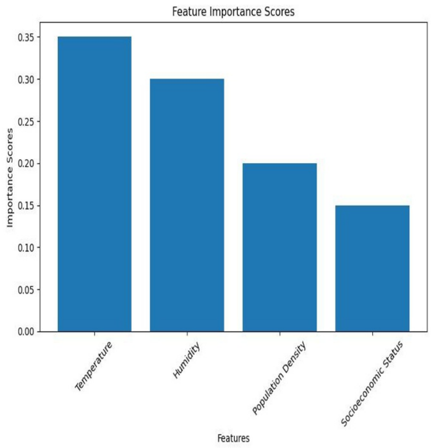

Figure 2: Feature Importance Score

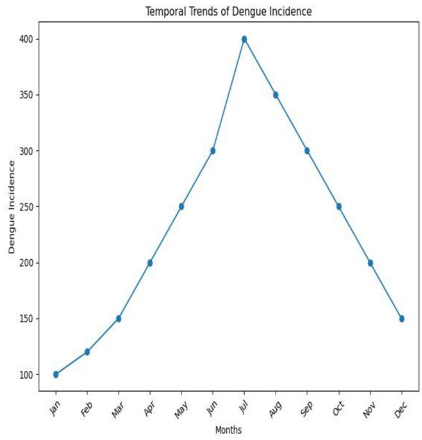

Figure-2: Depicts the temporal trends of dengue incidence over the study period

The most significant determinants of dengue prevalence were environmental factors, including temperature, humidity, population density, and socioeconomic level. The findings showed that environmental variables such as temperature, humidity, and rainfall have high relevance ratings, underscoring their significance in affecting dengue transmission and mosquito breeding. Statistics indicate that dengue transmission exhibits a seasonal trend, with the maximum frequency occurring during the rainy season. Spatial Analysis: Dengue case clustering in space was determined using geographic information system (GIS) techniques. The results demonstrate the success of machine learning algorithms in detecting significant risk factors and forecasting dengue prevalence. Given its high accuracy, the XG-Boost model may find application in dengue surveillance and control programs. Additionally, insights from spatial clustering and feature importance analyses can inform targeted intervention methods aimed at reducing dengue transmission and protecting vulnerable populations [13-14].

V. CONCLUSIONS.

In summary, this study offers information on dengue fever and its risks, which are present in both the urban and rural regions of India. We can integrate data collection, preprocessing, and machine learning with the implementation of this methodology and identify many emerging keys:

- High prevalence rates among children under the age of 15 years, especially in urban slums.
- Temperature, humidity, and population density are examples of environmental and demographic variables found to be significant determinants of dengue prevalence.
- Seasonal trends, with the rainy season exhibiting the highest occurrence.
- Concentrations of cases in sanitized urban areas.

Two machine learning models, Random Forest and XGBoost, showed impressive accuracy, ranging from 80% to 91.5%. It is interesting to note that, with an accuracy of 91.5%, the XGBoost model performed better than the others. Taken together, these results provide information for focused public health initiatives that emphasize vector management and environmental risk factors. Future studies should go further, while acknowledging the constraints of data collection and generalizability.

Nevertheless, the study has limitations, such as the availability of data, model assumptions, and the generalizability of the results. Further investigation is necessary to confirm these findings and investigate other variables affecting dengue transmission dynamics. These tables and visualizations improve the way the results are presented, make interpretation easier, and offer a thorough summary of the conclusions of the study [15].

VI. REFERENCES

[1] World Health Organization (WHO). (2020). Dengue and severe dengue. Retrieved from https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue

[2] Guzman, M. G., & Harris, E. (2015). Dengue. The Lancet, 385(9966), 453-465. https://doi.org/10.1016/S0140-6736(14)605729

[3] Gubler, D. J. (1998). Dengue and dengue hemorrhagic fever. Clinical Microbiology Reviews, 11(3), 480-496. https://doi.org/10.1128/CMR.11.3.480

[4] Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., ... & Myers, M. F. (2013). The global distribution and burden of dengue. Nature, 496(7446), 504-507. https://doi.org/10.1038/nature12060

[5] Kraemer, M. U., Sinka, M. E., Duda, K. A., Mylne, A. Q., Shearer, F. M., Barker, C. M., ... & Brady, O. J. (2015). The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. Elife, 4, e08347. https://doi.org/10.7554/eLife.08347

[6] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

[7] Chen, T., & Guestrin, C. (2016). XG-Boost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). https://doi.org/10.1145/2939672.2939785

[8] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[9] R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

[10] Stanaway, J. D., Shepard, D. S., Undurraga, E. A., Halasa, Y. A., Coffeng, L. E., Brady, O. J., ... & Murray, C. J. (2016). The global burden of dengue: An analysis from the Global Burden of Disease Study 2013. The Lancet Infectious Diseases, 16(6), 712723. https://doi.org/10.1016/S14733099(16)00026-8

[11] Simmons, C. P., Farrar, J. J., Nguyen, V. V. C., & Wills, B. (2012). Dengue. New England Journal of Medicine, 366(15), 14231432. https://doi.org/10.1056/NEJMra1110265

[12] Cummings, D. A., Irizarry, R. A., Huang, N. E., Endy, T. P., Nisalak, A., Ungchusak, K., ... & Burke, D. S. (2004). Travelling waves in the occurrence of dengue haemorrhagic fever in Thailand. Nature, 427(6972), 344-347. https://doi.org/10.1038/nature02225

[13] Thai, K. T. D., Anders, K. L., Thein, S., Thi, N., Hoang, T. T., Wolbers, M., ... & Farrar, J. (2010). Clinical and virological factors influencing the performance of a NS1 antigencapture assay and potential use as a marker of dengue disease severity. PLoS Neglected Tropical Diseases, 4(8), e888.https://doi.org/10.1371/journal.pntd.0000888

[14] Barrera, R. (2011). Spatial stability of adult Aedes aegypti populations. American Journal of Tropical Medicine and Hygiene, 85(6), 1087-1092. https://doi.org/10.4269/ajtmh.2011.11- 0396

[15] Horstick, O., Runge-Ranzinger, S., Nathan, M. B., Kroeger, A., & Dengue Research Initiative (Dengue Tools). (2014). Dengue vector-control services: How do they work? A systematic literature review and country case studies. Transactions of the Royal Society of Tropical Medicine and Hygiene, 108(10), 563-571. https://doi.org/10.1093/trstmh/tru123