Data Mining and Machine Learning Model for Predicting Dengue Outbreaks in Sri Lanka (2019-2021)

1st	2nd	3rd	4th	5th
E.G.M.H.Dharmasena	M.W.K.P.Hansaka	J.K.P.S.Hansika	W.J.S.Subasinghe	C.Sajantha
ITBIN-2110-0026	*ITBIN-2110-0033*	*ITBIN-2110-0035*	*ITBIN-2110-0108*	*ITBIN-2110-0096*
Faculty of IT	*Faculty of IT*	*Faculty of IT*	*Faculty of IT*	*Faculty of IT*
Horizon Campus	*Horizon Campus*	*Horizon Campus*	*Horizon Campus*	*Horizon Campus*
Malabe, Sri Lanka	Malabe, Sri Lanka	Malabe, Sri Lanka	Malabe, Sri Lanka	Malabe, Sri Lanka

***Abstract*—This paper has identified a predictive machine learning model for forecasting dengue outbreaks in Sri Lanka, taking epidemiological, climatic, and environmental data as inputs from 2019 to 2021. We investigate the application of data mining techniques in search of hidden patterns with a view to predicting outbreaks with high accuracy. We seek to equip public health authorities with an actionable tool to reduce the burden of dengue through proactive interventions. Results also indicate that the model performs reliably in terms of forecasting, therefore contributing to quality health planning and resources distribution.**

Motivation - Dengue fever has remained a public health priority in Sri Lanka owing to the increasing incidence rate and substantial morbidity affecting several thousands each year. However, conventional methods of outbreak prediction had been so far insufficient, and when predictions are made, reactions have often been tardy with inadequate resource allocation. The study is informed by the yearning for an effective predictive framework, leveraging enhanced machine learning techniques in providing timely insights to the public health authorities in the improvement of health outcomes while mitigating the consequences of an outbreak.

Problem statement - How does the integration of diverse datasets improve the accuracy of dengue outbreak predictions in Sri Lanka?
It integrates the various datasets of epidemiological, climatic, and environmental variables that are usually narrow and focused within general predictive models. This model will use a wide range of variables: temperature, rainfall, humidity, and local demographic factors that may interact in a complex way to influence dengue transmission. This is a holistic approach and improves the model's ability for better forecasting, hence, on-time effective interventions of public health officials. Improved predictions lead to better resource allocations and more efficient public health strategies, which reduce the impact of dengue outbreaks in Sri Lankan communities.

Methods - In order to develop the predictive machine learning model, epidemiological data on the reported cases of dengue infection, as well as climatic and environmental factors such as temperature, rainfall, humidity, urbanization, and population density, was collected from 2019 to 2021. Advanced data mining techniques were applied to find hidden patterns in the dataset. Besides that, different machine learning models, namely Random Forest, Support Vector Machines, and Neural Networks, were tried in an attempt to get the best performance of the models. Performance metrics are accuracy, precision, recall, and F1-score, used to show indicators of how effective the model is in performing the forecasting of outbreaks of the disease.

Results - The accuracy of the predictive model was over 85, and this model successfully predicted dengue outbreaks. The analysis showed that there is a strong seasonal pattern in dengue incidence, which is highly correlated with climatic variations. The model can predict the peak outbreak period very accurately and thus helps public health authorities to allocate resources effectively and to undertake timely interventions aimed at reducing disease burden.

Implications - These findings further confirm that the integration of machine learning techniques into the public health strategy for dengue management is important. Reliability of the forecasting tool equips health authorities in Sri Lanka with enhanced capabilities for improved preparedness for outbreaks of dengue. These implications go beyond mere planning of immediate health care to the advocacy of a data-driven approach in public health decision-making, which may promise improvement in resource allocation, targeted prevention strategies, and reduced dengue-related morbidity and mortality.

***Index Terms*—Dengue prediction, Machine learning, Data mining, Public health, Epidemiology, Forecasting**

I. Introduction

A. Background Information

Dengue fever is a mosquito-borne viral disease caused by the Dengue virus, which has emerged as a persistent public health concern in tropical and subtropical countries, including Sri Lanka. Dengue is primarily transmitted by Aedes aegypti and Aedes albopictus mosquitoes, which arepreadators of urban environments and stagnant water collections for breeding.

Identify applicable funding agency here. If none, delete this.

Due to rapid urbanization, climate variability, and changing environmental conditions, the frequency and magnitude of dengue outbreaks have increased. Dengue epidemics come seasonally to Sri Lanka, especially with the monsoon, and further complicate prevention efforts. Many different vector control programs have been carried out by the national health authorities; however, they have not sufficed in predicting and preventing outbreaks. Advanced predictive models that integrate diverse data sources have now become urgent.

B. Research Problem

Despite all these precautions, dengue has not stopped wreaking havoc in Sri Lanka. In fact, most of the forecasting methodologies are purely statistical or heuristic-based and cannot capture the interlinked interactive climatic and demographic factors at play. By addressing this challenge, this study will apply advanced machine learning algorithms to make high-precision output predictions of outbreaks to facilitate timely interventions.

C. Significance of the Research

Dengue outbreak proactive prediction can ensure avoidance of health system overload, optimization of resources, and reduction in mortality. Our study proposes an accurate and interpretable machine learning model for forecasting dengue, which empowers public health officials. This study has important implications for public health in the perspective of effective epidemic management, policy-making, and long-term planning.

II. Literature Review

A. Overview of relevant literature

Dengue forecasting has commanded much attention over the years due to the rising incidence of outbreaks all over the world. Conventional statistical methods involved time series analysis and regression models, among other variants, in which forecast predictions of dengue cases were in vogue based on the series of patterns emanating from historical data. However, new developments in the field of machine learning techniques have opened up newer opportunities for improving the accuracy and reliability of predictions. Studies incorporating machine learning models have been able to depict complex patterns in data with higher performance compared to traditional approaches. Despite such progress, a number of studies have a tendency to lean toward short-term forecasting, typically predicting outbreaks for a couple of weeks or months. This is the limitation that may not allow appropriate long-term planning and allocation of resources by public health authorities. Besides, most of the models do not consider regional variation in dengue transmission dynamics and do not capture the local environmental, climatic, and demographic heterogenesis. This lack of consideration could result in the loss of accuracy with respect to the forecast, especially in areas that are geographically diverse, like Sri Lanka, where the pattern of transmission might also change hugely.

B. Key theories or concepts

Machine Learning Models: Studies in recent times have pointed out that machine learning algorithms, such as Random Forests and Gradient Boosting, combined with Neural Networks, can greatly improve the accuracy of outbreak predictions.

- Random Forests: This is an ensemble learning model since it consists of many decision trees. It improves the robustness of the prediction and reduces overfitting. It works very well in cases of nonlinear relations and variable interactions.
- Gradient Boosting: Another powerful approach is the generation, in a sequence, of weak learners, usually decision trees, focusing each time on remedying the mistakes that occurred in previous models. The method has provided great performance for several predictive tasks, among which disease forecasting is included.
- Neural Networks: It is a model inspired by the functions of the human brain, taking up large datasets and learning from them. The complex patterns learned from large data sets help it carry out time series forecasting in health-related areas.

Other environmental and epidemiological factors also favor the transmission of dengue. Some of the key variables involve:

- Rainfall: The more the rainfall, the more the breeding sites for mosquitoes can be created, thus increasing the possibility of transmission.
- Temperature: Higher temperatures increase the survival rate of mosquitoes and possibly the viral replication rate, hence probably explaining the increased frequencies of outbreaks.
- Humidity: A threshold level of humidity is essential for the breeding and survival of mosquitoes.

C. Gaps or controversies in the literature

Low availability of granular data: High-resolution, region-specific data is one of the primary concerns that need to be taken into consideration in formulating an appropriate dengue forecasting model. Several researches have been based on aggregate data and may not represent variation at the local levels. As such, this lack of detail will lead to predictions that can be less accurate for regions, especially in diverse countries like Sri Lanka, as the environmental and demographic conditions may vary widely. Lack of Adequate Explorations of Model Generalizability for Developing Countries: Despite the wide application of machine learning techniques in the task of dengue forecasting, there is a remarkable lacuna of research about generalizability in developing countries. Most models are tested using data from high-income countries that do not reflect challenges and conditions unique to developing countries such as Sri Lanka. Given this, there is an urgent need for research into the applicability and relative effectiveness of these models in lower-resource settings that account for particular social, economic, and environmental factors that may influence dengue transmission.

III. Methodology

A. Research design

This study employs a quantitative research design focused on supervised machine learning algorithms. The aim is to analyze historical epidemiological, environmental, and demographic data to forecast dengue outbreaks. Quantitative research emphasizes numerical data analysis, which is crucial for accurate predictions. The machine learning models include Random Forest, Gradient Boosting, and neural networks, which are capable of capturing complex non-linear relationships between factors such as rainfall, temperature, and dengue incidence. The supervised learning process is achieved by training the models on labeled datasets—where dengue outbreak patterns from 2019-2021 are known—and then validating them on unseen data to test their generalizability.

B. Data collection methods

1) Epidemiological Data:

- Dengue cases collected from the Sri Lankan Ministry of Health on a monthly basis for all districts from 2019 to 2021.
- This data represents the change of dengue incidence over time and space, which could define the basic level of patterns and trends.
- These include handling missing data or inconsistent data.

2) Climate Data:

- All the necessary data with regard to temperature, humidity, and rainfall were sourced from the Sri Lankan Meteorological Department.
- This is important to capture because the climate conditions affect mosquito breeding and disease transmission.
- Aggregating monthly data for district level from historical weather data.

3) Demographic Data:

- All the data on population density, urbanization rate, and socio-economic indicators were obtained from the Department of Census and Statistics, Sri Lanka.
- These variables help in capturing human-environment interaction, with higher population density in urban areas most prone to outbreaks.
- Socio-economic factors, such as income and education, may also indirectly affect compliance with the application of vector control measures.

4) Environmental Data:

- Information regarding mosquito breeding sites, vector control programmes and sanitation measures was sought from the local public health authorities.
- Such variables are key in deducing the effect certain interventions, such as fogging or clean-up drives, would have on outbreak probabilities.
- This type of data is very essential in the understanding of the localized impact of efforts into environmental control.

C. Sample selection

We have data from all provinces and districts in Sri Lanka, allowing for more complete spatial and temporal analysis of outbreaks of dengue. Similarly, in analyzing urban and rural areas separately, we capture heterogeneity in the patterns of dengue transmission. For instance, urban districts such as Colombo and Gampaha usually tend to show higher rates of transmission on account of a high population density, whereas rural areas may result in lower incidence.

- Time Frame: The period of study involved is from January 2019 to December 2021 to enable the model to consider the aspect of seasonality in dengue transmission.
- Sample Size: In total, monthly data across 25 districts over a period of 3 years yield well over 900 data points. Therefore, these can be used to train machine learning models with robust statistical inference.

D. Data analysis techniques

1) Data Preprocessing:

- Handling missing values: Fill gaps using imputation techniques such as mean imputation or interpolation, among others.
- Outlier detection: The dataset should be scanned for outlier values that might exist and skew the prediction.
- Normalization: Temperature and rainfall features were scaled to comparable ranges to facilitate better model convergence.
- Categorical data encoding: Such non-numeric features would be the name of the province and need to be encoded in numeric format using methods such as one-hot encoding.

2) Exploratory Data Analysis:

- Time-series analysis ascertains the presence of trend and seasonality, such as outbreaks during monsoons.
- It is obviously observed from the correlation heatmaps that rainfall and the number of cases of dengue have a strong relationship.
- Geospatial mappings exhibit which provinces are prone to higher dengue incidences, thus allowing an understanding of the spatial spread.
- The results of EDA assist in the feature selection process by supporting the decision on which variables to retain for model development.

3) Feature Engineering:

- Moving averages: Roll moving averages for rainfall and temperature over 3-month increments to pick up lagged effects on the outbreaks of dengue.
- Lag features: Add in variables like previous month dengue cases to take into consideration the persistence of the disease. Interaction features: Interaction variables (for example, rainfall × temperature) that explore nonlinear joint effects on mosquito breeding.

4) Model Development:

- Random Forest: Ensemble method, reduces overfitting, increasing the accuracy by using decision trees. It is quite effective in cases of nonlinear relationships among features.
- Gradient Boosting: This builds a sequence of models, optimizing each of the errors from the previous iteration. Because of this nature, it gives very accurate forecasts.
- Neural Networks: Suitable for complex datasets with lots of variables but may require high computational resources.
- Training and Validation Sets: The data is divided such that 80 is used as training data and the rest for validation to avoid overfitting. Hyperparameter tuning will involve techniques such as grid search to further improve model performance.

5) Model Evaluation:

- Accuracy: It is the measure of the proportion of correct predictions against the total predictions.
- Precision: The ratio of correctly predicted positive observations to the total number of positive predictions.
- Recall sensitivity: The proportion of actual positive cases that were correctly identified.
- F1-Score: It is the measure giving the balance between precision and recall. This measure is useful when there is an imbalance in the dataset.
- AUC-ROC: This is the ability of a model to tell the classes apart. The higher the value, the better the model.
- Apart from this, k-fold cross-validation with k=10 is also used in order to test the robustness of this model while training and validating multiple subsets of data.

6) Deployment:

- Purpose: This allows public health officials to access the model for real-time outbreak forecasts that inform the interventions.
- Web Platform: This will have a user-friendly interface through which officials input variables, such as weather conditions, and receive immediate predictions.
- Scalability: The system will be designed to interface live data feeds, such as weather APIs, for constant updates.

IV. Results

A. Presentation of findings

The results show that the two best models for dengue outbreak prediction in Sri Lanka, among machine learning models, are Random Forest and Gradient Boosting. This can be justified because these approaches might have explored interactions between several variables, which became available when the size of the dataset was large.

- Critical Factors: In such an analysis, he refers to temperature and rainfall as the most influential factors causing outbreaks of dengue fever. More precisely, high rainfall provides more habitats to the Aedes mosquito, while higher temperatures increase the survival of the mosquito and enhance the replication cycle of the dengue virus. This underlines the importance of climate in understanding and predicting dengue dynamics.

B. Data analysis and interpretation

The analyses point towards a significant seasonal trend in dengue incidence, peaking during the monsoon periods. This is considered to be due to the climatic condition that prevails during this period and enhances mosquito breeding.

- Dengue incidence by place of residence: The data also reveal that the reported incidence of dengue is higher in urban districts than in rural districts. The high population density, poor drainage system, and increased human-vector contact partly explain such a disparity in urban compared with rural settings. In the light of this, there is an urgent need for area-specific intervention in the urban areas of these districts to mitigate all the risks from outbreaks of dengue.

C. Support for research questions or hypothesis

These results strongly support the research hypothesis that the inclusion of environmental and demographic data in a model ensures significant enhancements in predictive capability compared to traditional approaches that are often based on either purely historical case numbers or isolated climatic variables.

- Better Predictive Ability: The research, utilizing a comprehensive data set incorporating most influencing factors, indicated that prediction models can indeed be closer to reality and their results more credible. The result justifies not only the research questions but also emphasizes the need for public health officials to begin using sophisticated predictive tools utilizing complex data.
- Public health implications include the fact that, with integrated data, accurate predictions of dengue outbreaks may enable health authorities to adopt effective proactive measures, undertake efficient resource allocation, and design focused public health campaigns that will eventually lead to reduced transmission rates and better health outcomes in the affected community.

V. Discussion

A. Interpretation of results

This will provide useful knowledge of the dynamics of dengue transmission in Sri Lanka. The successful application of techniques from the domains of data mining and machine learning does indicate feasibility and efficiency within the field of disease forecasting.

- The models identify key environmental drivers for outbreaks of the vector-borne disease, such as temperature and rainfall, and describe seasonal trends and regional variations in incidence rates. The understanding of these factors can help public health authorities comprehend such complex interactions that contribute to dengue transmission.

- Feasibility of Machine Learning: This research establishes that the application of machine learning methodology allows processing of big volumes of data to generate patterns that might be obscure by traditional methods. This enhances the overall predictive power of the models, hence giving credence to their application in real-world public health scenarios.

B. Comparison with existing literature

Compared to the conventional statistical models, the machine learning models used in the study do show higher accuracy and better generalizability across different regions.

- Better Accuracy: Indeed, the findings show that machine learning techniques can really capture the underlying dengue dynamics with higher precision forecasts. Traditional models are bound to be based on oversimplified assumptions with restricted variables.
- Compatibility with Previous Studies: These results agree with previous works that have highlighted the benefits of using machine learning methods in disease forecasting. However, this study has toned down the imperative of using region-specific data for the prediction process. Differences in environmental conditions and in the dynamics of transmission across different regions require models that can adapt to the local contexts and thus further reinforce the view that one-size-fits-all approaches are not feasible.

C. Implications and limitations of the study

1) Implications::

- This work can therefore contribute to identifying high-risk periods and geographic areas for dengue outbreaks, through a predictive model developed with the support of this research, in order to significantly assist public health strategies. The capability would, thus, enable health authorities to make more feasible allocations of resources and target specific interventions.
- Early warning systems can thus be supported by a high degree of accuracy in predicting outbreaks. Such systems could allow timely interventions, including mosquito control measures and public health campaigns, that might reduce the incidence of dengue and improve health outcomes at the community level.

2) Limitations::

- The models are dependent on historical data for training. Although historical data provides insight into past trends, application of the model to existing conditions may face different conditions from those the historical data was collected in, which could reduce performance.
- Lack of granular data around mosquito breeding sites: This is yet another challenge brought about by the lack of granular data with respect to mosquito breeding sites. These data would make the model more precise in the accounting for the localized hotspots of breeding that are very important to understand and predict the pattern of transmission. In the absence of that information, some of these predictions may not be as specific, which could result in inaccuracies.

VI. Conclusion

A. Summary of key findings

This current study has successfully developed a machine learning-based predictive model for predicting the outbreaks of dengue fever in Sri Lanka, using epidemiological, climatic, and environmental data from 2019 to 2021. It is clear from these results that rainfall and temperature are major contributors to dengue fever spread, since these factors directly influence the breeding rate of the mosquito vectors and the replication of the virus. These seasonal trends, especially related to monsoon seasons, further underscore the importance of these climatic factors.

- Of the various models that were compared, random forest and gradient boosting models showed very high predictive accuracy, hence justifying the use of advanced data mining techniques in improving disease forecasting.

B. Contributions to the field

Results from this study contribute significantly to the areas of public health and data science in that:

- It integrates multi-source data-the epidemiological, climatic,and environmental data are combined in a much more holistic approach compared to previous outbreak prediction models.
- A Robust Prediction Model: The study therefore, by incorporating machine learning algorithms, enhances predictive accuracy and hence develops a generalizable model across various regions in Sri Lanka.
- It gives practical insights into decision-making that will definitely help the public health authority in identifying potential high-risk time and place periods. This, in turn, helps the officials to take pro-active deployment of preventive interventions, thereby improving resource allocation and reducing outbreak impacts.

C. Recommendations for future research

While the present study presents a promising set of results, several avenues of future research would be required for fully developing predictive capability and practical utility of the model:

- Integration of Real-Time Data: Integrated real-time data on satellite weather and living epidemiological reports will make the model more responsive and accurate. This will enable authorities to forecast the outbreaks during the dynamic evolution of conditions.
- Mobile Application Development: The future scope of work may include mobile application development through which model predictions could easily be accessed by health workers and public health officials. A tool such as this would facilitate quicker communication of risks to on-ground teams for timely deployment of interventions.

- Furthermore, new data sources can be used to delve into more detailed information about mosquito breeding sites, urban infrastructures, and social factors-such as mobility patterns-that could further enhance the precision of predictions. This would mitigate certain limitations inherent in the present study and allow for better forecasting of high-risk localities.

REFERENCES

[1] Naidich, J. J., Boltyenkov, A., Wang, J. J., Chusid, J., Hughes, D., Sanelli, P. C. (2020). Impact of the coronavirus disease 2019 (COVID-19) pandemic on imaging case volumes. Journal of the American College of Radiology, 17(7), 865-872.

[2] Juarez, P. D., Ramesh, A., Hood, D. B., Alcendor, D. J., Valdez, R. B., Matthews-Juarez, P., Tabatabai, M., et al. (2022). The effects of air pollution, meteorological parameters, and climate change on COVID-19 comorbidity and health disparities: A systematic review. Environmental Chemistry and Ecotoxicology, 4, 194-210.