Research Methodology and Scientific Writing

CST 395-2

Group No: 24

Dengue Outbreak Risk Prediction Models for Sri Lanka Using Machine Learning-based Approach

Computer Science and Technology

Department of Computer Science and Informatics

Uva Wellassa University

June 2025

Dengue Outbreak Risk Prediction Models for Sri Lanka Using Machine Learning-based Approach

1. Introduction

Dengue remains a major public health concern in Sri Lanka, with outbreaks exhibiting strong associations with meteorological patterns and socio-environmental factors. Traditional forecasting methods (e.g., ARIMA, ETS) have been useful but often lack sufficient granularity and adaptability. Recent advances in neural network approaches, particularly RNN-based models such as LSTM, demonstrate superior performance in capturing temporal dependencies and forecasting vector-borne diseases [1], [2]. However, no district-level, weekly, end-to-end pipeline utilizing LSTM has yet been developed for Sri Lanka—a gap this research aims to address.

2. Literature Review

Dengue fever is a mosquito-borne viral disease endemic to tropical regions, including Sri Lanka. Forecasting its outbreaks is critical for public health resource allocation. Numerous studies have explored the use of climatic and epidemiological data to predict dengue incidences using time-series models or basic machine learning methods. However, most lack weekly granularity, integration of socio-environmental variables, or real-time update pipelines.

Withanage et al. developed a forecasting model for the Gampaha district using lagged climatic variables, emphasizing rainfall and humidity [3]. However, the study used monthly data and linear regression methods, limiting its responsiveness to sudden outbreak surges. **Madushani and Talagala** explored hierarchical time-series models for dengue forecasting across administrative levels (country → province → district), improving coherence but still relying on statistical methods like ETS and ARIMA [2].

Globally, **Kakarla et al.** used ensemble ML models in India with weather and demographic data, showing increased accuracy over traditional statistical models [4]. However, their model focused on monthly rather than weekly predictions and was not contextualized to Sri Lanka. **Saleh and Baiwei** demonstrated the power of deep learning, particularly LSTM models, to forecast dengue with higher temporal precision [5]. Their study, however, was experimental and lacked public health deployment focus.

Recent work by **Mujeed et al.** shows that attention-augmented LSTM (A-LSTM) outperforms both classical and deep models in dengue prediction due to its ability to focus on critical time points in multivariate sequences [6]. Additionally, **Li et al.** proposed spatially aware RNN models integrating climatic and location data, demonstrating strong potential for geographically distributed disease modeling [7].

While the studies confirm the growing adoption of ML/DL for outbreak prediction, their implementation in **resource-constrained**, **district-level**, and **real-time** public health systems—especially in **Sri Lanka**—is largely unaddressed. Furthermore, few of these models feature **cloud-based pipelines** or **decision-maker-oriented dashboards** to enhance their practical use.

3. Identified Research Gap:

Despite the demonstrated potential of machine learning—particularly RNN-based approaches like LSTM and A-LSTM—in predicting dengue outbreaks:

- No existing model integrates **weekly epidemiological, meteorological, and socio-environmental data** at **district granularity** in Sri Lanka.
- There is a **contextual gap**: most studies are international or national-level models, not tailored to **local Sri Lankan needs**.
- A **methodological gap** exists most local studies rely on classical models (ARIMA, regression), while RNN-based architectures remain underexplored.

4. Research Aim, Questions & Objectives

Aim:
To develop a district-level machine learning model for predicting dengue outbreak risks in Sri Lanka by integrating weekly epidemiological, meteorological, and socio-environmental data streams.

Research Questions:

1. Which epidemiological, meteorological, and socio-environmental variables show statistically significant correlations with dengue outbreak patterns across Sri Lankan districts?
2. How can the ML model be optimized to process weekly surveillance data for district-level outbreak forecasting, and which algorithm (e.g., LSTM) is most suitable?
3. What infrastructure is required to enable automated weekly model updates in resource-constrained settings?
4. How can an interactive risk visualization dashboard be designed to effectively communicate district-level dengue outbreak forecasts to public health stakeholders?

Objectives:

1. To conduct a systematic review of dengue outbreak predictors in tropical climates.
2. To preprocess district-level dengue, meteorological, and socio-environmental datasets.
3. To design an LSTM-based architecture for temporal pattern recognition.
4. To develop a cloud-based data ingestion pipeline for weekly forecasts.
5. To build a prototype interactive risk visualization dashboard.

5. Conceptual Framework

The conceptual framework outlines how different data inputs, analytical processes, and technological components interact to achieve the research objectives. It establishes the logical flow from data collection to decision-support, driven by the capabilities of Recurrent Neural Network (RNN)-based models such as LSTM.

Key Components and Relationships

1. **Input Variables**
   - **Epidemiological Data**: Weekly district-level dengue case counts
   - **Meteorological Data**: Temperature, rainfall, humidity, wind speed
   - **Socio-environmental Data**: Population density, land use, waste disposal patterns, vector surveillance indices
2. **Data Preprocessing**
   - Feature engineering (e.g., lag features, seasonality)
   - Normalization and handling missing data
3. **Machine Learning Component**
   - **Model Type**: RNN-based (LSTM/A-LSTM) architecture
   - Designed for temporal sequence learning and forecasting
   - Optimized with grid search or other tuning methods
4. **Infrastructure**
   - **Cloud-hosted Pipeline**: For data ingestion, model training, and weekly auto-updates
   - **APIs**: For real-time integration with data sources and visualization modules
5. **Output**
   - **Forecasted Weekly Risk Levels**: For each district
   - **Risk Classification**: (e.g., Low, Medium, High)
6. **User Interface**
   - **Interactive Visualization Dashboard**: Map-based, tabular and graphical risk views
   - Usability testing with public health officials
7. **Feedback Loop**
   - Health officials' feedback used to improve model interpretability and dashboard functionality

6. References

[1] K. Roster and F. A. Rodrigues, “Neural Networks for Dengue Prediction: A Systematic Review,” *arXiv*, Jun. 2021. [Online]. Available: https://arxiv.org/abs/2106.12905

[2] L. S. Madushani and T. S. Talagala, “Hierarchical Forecasting of Dengue Incidence in Sri Lanka,” *arXiv*, Dec. 2021. [Online]. Available: https://arxiv.org/abs/2112.00992

[3] G. P. Withanage *et al.*, “A forecasting model for dengue incidence in the District of Gampaha, Sri Lanka,” *Parasites & Vectors*, vol. 11, p. 262, 2018. [Online]. Available: https://doi.org/10.1186/s13071-018-2828-2

[4] S. G. Kakarla *et al.*, “Weather integrated multiple machine learning models for prediction of dengue prevalence in India,” *Int. J. Biometeorol.*, vol. 67, pp. 285–297, Feb. 2023. [Online]. Available: https://doi.org/10.1007/s00484-022-02405-z

[5] A. Y. Saleh and L. Baiwei, “Dengue prediction using deep learning with long short-term memory,” *Proc. eSmarTA*, pp. 1–5, 2021. [Online]. Available: https://doi.org/10.1109/eSmarTA52612.2021.9515734

[6] M. A. Mujeed *et al.*, “Prediction of dengue cases using the attention-based long short-term memory (LSTM) approach,” *Geo. Health*, [Online]. Available: https://geospatialhealth.net/gh/article/view/1176

[7] Z. Li, J. Xin, and G. Zhou, “An integrated recurrent neural network and regression model with spatial and climatic couplings for vector-borne disease dynamics,” *arXiv*, Jan. 2022. [Online]. Available: https://arxiv.org/abs/2201.09394