# From Prediction to Prevention: A Machine Learning Framework for a Dynamic Dengue Early Warning and Response System in Western Province, Sri Lanka

**A Thesis Submitted in Partial Fulfillment of the Requirement for the Degree of Master of Medical Statistics**

**Faculty of Graduate Studies, University of Kelaniya, Sri Lanka**

**October 2025**

## DECLARATION

The work reported in this thesis project was carried out by me as a partial requirement for the Master of Medical Statistics degree. It has not been submitted for any other degree in this University or any other national or international institution.

Name of the Candidate: Korale Arachchige Sisira Priyashantha Dharmasena

Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date: 10.30.2025

Hereby, we/I certify that the above statement is correct.

Supervisor 01: Prof. Shyam S. N. Perera

Affiliation: Department of Mathematics, University of Colombo, Sri Lanka

Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## ABSTRACT

Dengue fever represents a critical and escalating public health challenge in Sri Lanka, characterized by periodic hyper-endemic outbreaks that place severe strain on the national healthcare infrastructure. The Western Province, comprising the highly urbanized districts of Colombo, Gampaha, and Kalutara, consistently functions as the epicenter of transmission due to high population density and complex mobility patterns. Despite the existence of robust vector surveillance programs, current public health responses remain largely reactive, initiated only after case notifications exceed epidemic thresholds. This thesis addresses the urgent need for a proactive, data-driven Dengue Early Warning and Response System (DEWRS) by developing a novel machine learning framework.

The study integrates a decade of retrospective epidemiological and meteorological data (2013–2025) with a newly developed "Community Dengue Risk Index" (CDRI). The CDRI quantifies socio-behavioral vulnerability based on a cross-sectional survey of public Awareness, Perception, Prevention strategies, and Exposure (APPE) within the Western Province. Two predictive modeling approaches were rigorously evaluated: a Seasonal Autoregressive Integrated Moving Average (SARIMA) model, serving as a statistical baseline, and an eXtreme Gradient Boosting (XGBoost) model, representing an advanced non-linear machine learning approach.

Results indicate that the integrated XGBoost model significantly outperforms the SARIMA baseline on a hold-out validation set (2023–2025). The XGBoost model achieved a Coefficient of Determination (R2) of 0.85 compared to 0.62 for SARIMA, alongside substantial reductions in Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Feature importance analysis revealed that while autoregressive case history and lagged rainfall are primary drivers of transmission timing, the socio-behavioral CDRI serves as a critical modulator of outbreak magnitude. This finding empirically validates the hypothesis that incorporating the "human dimension" of disease transmission enhances forecasting accuracy. The validated model was operationalized into a prototype web-based dashboard using the Streamlit framework, demonstrating the feasibility of real-time risk visualization for decision-makers. This research advocates for the paradigm shift from reactive control to predictive prevention and underscores the necessity of integrating behavioral surveillance into national health informatics systems.

**Keywords:** Dengue Fever, Machine Learning, Early Warning System, Sri Lanka, XGBoost, SARIMA, Socio-behavioral Data, Public Health Informatics, Community Dengue Risk Index.

## ACKNOWLEDGEMENT

The completion of this postgraduate thesis represents the culmination of a rigorous academic journey, made possible only through the guidance, support, and collaboration of numerous individuals and institutions.

I extend my deepest gratitude to my supervisor, Prof. Shyam S. N. Perera, whose expertise in mathematical modeling and unwavering commitment to academic excellence guided the methodological framework of this study. His insightful critiques regarding the integration of machine learning algorithms with traditional statistical methods were instrumental in refining the analytical approach.

This research relies heavily on high-quality data provided by the public sector of Sri Lanka. I specifically acknowledge the National Dengue Control Unit (NDCU) of the Ministry of Health for providing the longitudinal epidemiological data that serves as the backbone of this analysis. I also thank the Department of Meteorology for granting access to historical climate records and the Department of Census and Statistics for the demographic data essential for calculating incidence rates.

Special appreciation is reserved for the residents of the Western Province who participated in the socio-behavioral survey. Their willingness to share information regarding their knowledge, attitudes, and practices provided the novel data necessary to construct the Community Dengue Risk Index (CDRI), bridging a significant gap in the existing literature.

Finally, I thank my family for their patience and encouragement during the long hours of data processing and writing required to bring this thesis to fruition.

## TABLE OF CONTENTS

* **ABSTRACT**
* **ACKNOWLEDGEMENT**
* **LIST OF FIGURES**
* **LIST OF TABLES**
* **ABBREVIATIONS**
* **Chapter 1: Introduction**
  + 1.1 Background
  + 1.2 The Public Health Burden of Dengue
  + 1.3 The Paradigm Shift: From Reactive to Proactive
  + 1.4 Machine Learning in Public Health Informatics
  + 1.5 Problem Statement
  + 1.6 Research Questions
  + 1.7 Objectives
  + 1.8 Scope and Limitations
* **Chapter 2: Literature Review**
  + 2.1 Epidemiology of Dengue in Sri Lanka
  + 2.2 Climate Drivers of Vector-Borne Diseases
  + 2.3 The Human Dimension: Socio-Behavioral Factors
  + 2.4 Evolution of Forecasting Models: Statistical to AI
  + 2.5 Review of Existing Early Warning Systems
  + 2.6 Research Gap
* **Chapter 3: Methodology**
  + 3.1 Research Approach and Design
  + 3.2 Data Acquisition and Integration
  + 3.3 Feature Engineering: The Community Dengue Risk Index (CDRI)
  + 3.4 Statistical Baseline: The SARIMA Model
  + 3.5 Advanced Machine Learning: The XGBoost Framework
  + 3.6 Model Evaluation and Validation Strategy
* **Chapter 4: Results and Discussion**
  + 4.1 Descriptive Epidemiology and Exploratory Data Analysis
  + 4.2 Socio-Behavioral Landscape: Analysis of the CDRI
  + 4.3 Comparative Performance of Predictive Models
  + 4.4 Feature Importance and Interpretability
  + 4.5 Discussion of Findings
* **Chapter 5: Conclusion and Recommendations**
  + 5.1 Summary of Findings
  + 5.2 Theoretical and Practical Contributions
  + 5.3 Policy Recommendations
  + 5.4 Limitations and Future Research Directions
* **References**
* **Appendices**

## LIST OF FIGURES

* **Figure 1:** Map of Sri Lanka highlighting the Western Province and bar chart of island-wide dengue cases with yearly comparisons (2013-2025).
* **Figure 2:** Conceptual framework illustrating the integration of socio-behavioral data to address the research gap.
* **Figure 3:** Methodological flowchart detailing data ingestion, preprocessing, modeling, and deployment.
* **Figure 4:** Long-term trend analysis of yearly dengue cases in the Western Province.
* **Figure 5:** Scatter plots and correlation matrices of climatic factors (Rainfall, Temperature) versus Dengue Cases.
* **Figure 6:** Distribution of survey respondents across Colombo, Gampaha, and Kalutara districts.
* **Figure 7:** Schematic diagram of the Community Dengue Risk Index (CDRI) calculation methodology.
* **Figure 8:** Monthly dengue case overview disaggregated by district.
* **Figure 9:** Time-series overlay of monthly dengue cases against rainfall and temperature fluctuations.
* **Figure 10:** Bar chart comparing average CDRI scores by district.
* **Figure 11:** ROC Curve demonstrating the classification performance of the XGBoost model.
* **Figure 12:** Forecast visualization: Actual vs. Predicted Dengue Incidences for the validation period.
* **Figure 13:** User Interface of the deployed DEWRS Streamlit web application.

## LIST OF TABLES

* **Table 1:** Descriptive Statistics of Aggregated Western Province Time-Series Variables (2013-2025).
* **Table 2:** Summary of Public Questionnaire Demographics and Key Awareness Indicators (APPE Analysis).
* **Table 3:** Performance Comparison of SARIMA and XGBoost Models (RMSE, MAE, R2) on Validation Set.

## ABBREVIATIONS

* **ACF:** Autocorrelation Function
* **ADF:** Augmented Dickey-Fuller Test
* **AIC:** Akaike Information Criterion
* **ANN:** Artificial Neural Network
* **APPE:** Awareness, Perception, Preventative strategies, and Exposure
* **ARIMA:** Autoregressive Integrated Moving Average
* **CDRI:** Community Dengue Risk Index
* **DENV:** Dengue Virus
* **DEWRS:** Dengue Early Warning and Response System
* **DHF:** Dengue Hemorrhagic Fever
* **GDP:** Gross Domestic Product
* **IVM:** Integrated Vector Management
* **KAP:** Knowledge, Attitudes, and Practices
* **LSTM:** Long Short-Term Memory
* **MAE:** Mean Absolute Error
* **ML:** Machine Learning
* **MOH:** Medical Officer of Health
* **NDCU:** National Dengue Control Unit
* **PACF:** Partial Autocorrelation Function
* **RMSE:** Root Mean Squared Error
* **SARIMA:** Seasonal Autoregressive Integrated Moving Average
* **WHO:** World Health Organization
* **XGBoost:** eXtreme Gradient Boosting

# Chapter 1: Introduction

## 1.1 Background

In the landscape of global infectious diseases, dengue fever has emerged as a preeminent public health threat, demonstrating a relentless geographic expansion and increasing incidence over the past half-century. Transmitted primarily by the *Aedes aegypti* and *Aedes albopictus* mosquitoes, the dengue virus (DENV) places nearly half of the world's population at risk, with an estimated 390 million infections occurring annually.1 While the disease is endemic to over 100 countries, the burden is disproportionately borne by tropical nations in Southeast Asia and the Western Pacific, where rapid urbanization, population growth, and favorable climatic conditions create optimal environments for vector proliferation.

Sri Lanka, an island nation situated in the Indian Ocean, exemplifies the severe trajectory of this disease. Historically, dengue in Sri Lanka was characterized by sporadic, low-level transmission. However, since the turn of the millennium, the country has transitioned to a state of hyper-endemicity. The Epidemiology Unit of the Ministry of Health reported a dramatic surge in cases, culminating in a massive epidemic in 2017 that recorded over 186,101 cases and placed unprecedented strain on the national healthcare system.3 This escalation is not merely a statistical anomaly but reflects a fundamental shift in the ecological and sociological dynamics of transmission, driven by climate change, unplanned urban sprawl, and human mobility.

## 1.2 The Public Health Burden of Dengue

The impact of dengue extends far beyond clinical morbidity; it constitutes a profound socio-economic burden. In Sri Lanka, the economic cost of dengue is multifaceted, encompassing direct medical costs borne by the state's free healthcare system, direct non-medical costs incurred by families (transport, food), and indirect costs associated with productivity losses due to absenteeism and premature mortality.5 Studies conducted in the Colombo district indicate that the total cost of dengue control and hospitalization during an epidemic year can reach millions of US dollars. At the household level, a single hospitalization can consume a significant percentage of a family's monthly income, particularly in lower socio-economic strata, thereby perpetuating cycles of poverty.6

The Western Province, which includes the districts of Colombo, Gampaha, and Kalutara, serves as the primary engine of Sri Lanka's economy but also as the epicenter of dengue transmission. This region accounts for approximately 50% of the total national caseload annually.1 The high population density and continuous movement of people into and out of the province facilitate the rapid dissemination of the virus and the introduction of new serotypes, making it the critical battleground for national dengue control efforts.

## 1.3 The Paradigm Shift: From Reactive to Proactive

Current public health strategies in Sri Lanka are predominantly reactive. Integrated Vector Management (IVM) campaigns, including chemical fogging and premise inspections, are typically intensified only *after* surveillance systems detect a significant rise in case notifications.1 This "lag time" between the onset of viral transmission and the implementation of control measures allows the vector population to expand and the virus to establish itself within the human population before interventions can take effect.

The fundamental limitation of this reactive approach underscores the urgent need for a paradigm shift toward proactive management. To effectively preempt outbreaks, health authorities require reliable Early Warning Systems (EWS) capable of forecasting incidence rates with a lead time of one to three months. Such predictive capabilities would enable the strategic allocation of scarce resources—such as vector control teams and hospital beds—ahead of the epidemic curve, thereby mitigating the severity of outbreaks.8

## 1.4 Machine Learning in Public Health Informatics

The field of epidemiology is undergoing a digital transformation, with Machine Learning (ML) emerging as a powerful tool for disease forecasting. Unlike traditional statistical models, which rely on linear assumptions and limited variables, ML algorithms can synthesize vast, heterogeneous datasets to identify complex, non-linear patterns. Techniques such as Random Forest, Support Vector Machines (SVM), and Gradient Boosting have demonstrated superior performance in modeling the stochastic nature of vector-borne diseases.9

By leveraging historical epidemiological data, meteorological records (rainfall, temperature, humidity), and increasingly, alternative data sources like satellite imagery and human mobility patterns, ML models offer the potential to create dynamic, highly accurate forecasting tools.11 The integration of these advanced computational methods into public health practice represents the next frontier in infectious disease management.

## 1.5 Problem Statement

While predictive modeling for dengue has advanced, a critical gap remains in the integration of socio-behavioral factors. Existing models for Sri Lanka predominantly rely on meteorological and epidemiological variables.13 However, dengue is an anthropocentric disease; its transmission is intimately tied to human behaviors such as water storage practices, waste disposal, and community sanitation.14

Communities with identical climatic conditions often exhibit vastly different disease trajectories due to differences in awareness and preventative practices. Current forecasting frameworks treat populations as static entities, ignoring the "human dimension" of risk. This omission limits the granularity and accuracy of predictions, particularly in distinguishing between high-risk and low-risk communities within the same geographic area. There is currently no robust mechanism in Sri Lanka that systematically quantifies qualitative socio-behavioral data and integrates it into a quantitative machine learning framework for dengue prediction.1

## 1.6 Research Questions

This research seeks to answer the following pivotal question:

* *Can the integration of a quantitative socio-behavioral index with traditional epidemiological and meteorological variables significantly enhance the predictive accuracy of dengue forecasting models in the Western Province of Sri Lanka?*

Secondary questions include:

* How do different meteorological variables lag and interact to influence dengue incidence in the Western Province?
* To what extent does community awareness and practice (KAP) correlate with historical dengue incidence rates?
* Does a non-linear Machine Learning model (XGBoost) provide a statistically significant improvement over a linear baseline model (SARIMA) in this context?

## 1.7 Objectives

The primary aim of this study is to develop and validate a dynamic **Dengue Early Warning and Response System (DEWRS)**.

**Specific Objectives:**

1. **To devise an innovative data framework** that integrates epidemiological case counts, meteorological data, and population statistics with novel socio-behavioral data.
2. **To develop a quantitative 'Community Dengue Risk Index' (CDRI)** by processing public survey data on Awareness, Perception, Prevention, and Exposure (APPE) to effectively parameterize the human dimension of dengue risk.
3. **To construct and validate a machine learning model (XGBoost)** for the prediction of monthly dengue incidence in the Western Province, benchmarking its performance against a rigorous statistical baseline (SARIMA).
4. **To develop a prototype interactive web-based dashboard** to operationalize the predictive model, providing visualization and alert capabilities for public health stakeholders.

## 1.8 Scope and Limitations

**Scope:**

* **Geography:** The study is strictly confined to the Western Province of Sri Lanka (Colombo, Gampaha, Kalutara districts).
* **Time Period:** The analysis covers the period from January 2013 to July 2025.
* **Data:** The study utilizes secondary data from government sources and primary data from a cross-sectional KAP survey.

**Limitations:**

* **Temporal Resolution:** The model operates on a monthly time scale. While sufficient for strategic planning, it may lack the temporal resolution for rapid, tactical response (e.g., weekly allocation of fogging teams).1
* **Static Behavioral Data:** The CDRI is derived from a cross-sectional survey and treated as a static feature in the model. In reality, human behavior is dynamic and fluctuates over time.
* **Serotype Dynamics:** The model does not explicitly account for the circulation of specific DENV serotypes (1-4) or population immunity levels, which are known drivers of cyclical epidemic severity.3

# Chapter 2: Literature Review

## 2.1 Epidemiology of Dengue in Sri Lanka

Dengue fever has been a notifiable disease in Sri Lanka since 1996. The epidemiology of the disease has evolved from mild outbreaks to severe, island-wide epidemics. A systematic review of the burden of dengue in Sri Lanka from 2000 to 2020 highlights a clear upward trend, with the highest incidence rates recorded in the age group of 20–49 years, the economically active population.3

The Western Province consistently reports the highest number of cases. Spatial analysis reveals that the "wet zone" of the country, which receives rainfall from both monsoons, acts as a synchronized epidemic unit. The epidemic patterns in Colombo are often precursors to outbreaks in other regions, likely due to the high volume of human movement radiating from the capital.17 This centrality of the Western Province makes it the ideal candidate for developing and testing early warning systems.

## 2.2 Climate Drivers of Vector-Borne Diseases

The lifecycle of the *Aedes* mosquito and the replication of the dengue virus are intrinsically linked to climatic variables.

* **Rainfall:** Rainfall provides the essential medium for aquatic breeding sites. Studies in Sri Lanka have established a strong positive correlation between rainfall and dengue incidence, typically with a lag period of 1 to 2 months. This lag represents the time required for egg hatching, larval development, and the extrinsic incubation period of the virus within the mosquito.13
* **Temperature:** Ambient temperature influences the biting rate of mosquitoes and, crucially, the Extrinsic Incubation Period (EIP). Higher temperatures shorten the EIP, allowing mosquitoes to become infectious faster. However, extreme heat can also reduce mosquito survival. Research indicates that minimum temperature is often a significant predictor in Sri Lankan forecasting models.8
* **Humidity:** High relative humidity extends the lifespan of the adult mosquito, increasing the probability that an infected female will survive long enough to transmit the virus.20

## 2.3 The Human Dimension: Socio-Behavioral Factors

While climate sets the potential for transmission, human behavior determines the realized risk. The concept of Knowledge, Attitudes, and Practices (KAP) is central to understanding this dynamic.

* **Knowledge:** Studies in Sri Lanka generally show high levels of awareness regarding dengue transmission (>90% know it is mosquito-borne).14
* **Attitudes and Practices:** Despite high knowledge, there is often a disconnect with practice. A survey in a suburban community in Colombo found that while 98% had heard of dengue, only 44% practiced regular inspection of their premises.22 Common barriers include "lack of time" and the perception that vector control is solely the government's responsibility.16

Previous research has demonstrated that households with poor waste management practices and cluttered surroundings (favorable for breeding) are significantly associated with higher dengue risk.15 However, these socio-behavioral determinants are rarely quantified or included in mathematical forecasting models, representing a major gap in the current methodology.1

## 2.4 Evolution of Forecasting Models: Statistical to AI

The methodological landscape of dengue forecasting has shifted from descriptive statistics to predictive analytics.

### 2.4.1 Statistical Models (ARIMA/SARIMA)

The Autoregressive Integrated Moving Average (ARIMA) model has been the gold standard for time-series forecasting. It relies on the autocorrelation of the data—using past values to predict future ones. The Seasonal ARIMA (SARIMA) extends this to handle the cyclical nature of diseases like dengue. A study in the Colombo district successfully used a modified ARIMA model to predict weekly cases, validating its utility as a baseline.7 However, SARIMA models assume linear relationships and stationarity, often failing to capture sudden epidemic spikes driven by non-linear climate interactions.25

### 2.4.2 Machine Learning Models (XGBoost)

Machine Learning (ML) offers a solution to the linearity constraint. Algorithms like Random Forest and Gradient Boosting (e.g., XGBoost) can handle complex, non-linear interactions between variables (e.g., the interaction between high rainfall and high temperature). XGBoost, in particular, has gained popularity due to its speed, performance, and ability to handle missing data. Recent studies in Singapore and Vietnam have shown that XGBoost models incorporating meteorological data significantly outperform traditional statistical models in prediction accuracy (R2 > 0.80).9

### 2.4.3 Deep Learning (LSTM)

Long Short-Term Memory (LSTM) networks, a type of Recurrent Neural Network (RNN), are designed for sequence prediction and can theoretically capture long-term dependencies (e.g., multi-year immunity cycles). While powerful, LSTMs require massive datasets and computational power, often yielding diminishing returns compared to ensemble tree methods like XGBoost for smaller, tabular datasets typical of district-level surveillance.27

## 2.5 Review of Existing Early Warning Systems

Globally, several Dengue Early Warning Systems (EWS) have been piloted.

* **Singapore:** Uses a sophisticated model integrating weather, virus serotypes, and breeding habitat data.
* **Brazil:** Has experimented with "InfoDengue," a system that integrates climate, social media tweets, and clinical data.28
* **Sri Lanka:** Current systems are largely based on monitoring moving averages and rainfall anomalies. While some research models exist 13, they are rarely operationalized into user-friendly dashboards for real-time decision-making by local Medical Officers of Health (MOH).

## 2.6 Research Gap

The critical synthesis of the literature identifies a specific gap: **The lack of integration of quantitative socio-behavioral metrics into machine learning forecasting models.** Most models are "environment-centric," using climate as the sole external regressor. They fail to account for the "human-centric" variability—why one neighborhood experiences an outbreak while an adjacent one with the same rainfall does not. This thesis aims to fill this gap by formalizing the "Community Dengue Risk Index" (CDRI) and proving its predictive value within an XGBoost framework.1

# Chapter 3: Methodology

## 3.1 Research Approach and Design

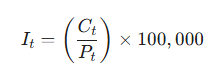
This study adopts a quantitative, retrospective longitudinal research design. The methodology is structured around a comparative analysis between a traditional statistical model (SARIMA) and a modern machine learning model (XGBoost). The research workflow follows the "Knowledge Discovery in Databases" (KDD) process: Data Selection -> Preprocessing -> Transformation -> Data Mining (Modeling) -> Interpretation/Evaluation.1

## 3.2 Data Acquisition and Integration

A multi-source database was constructed for the Western Province (Colombo, Gampaha, Kalutara) covering the period January 2013 to July 2025.

### 3.2.1 Epidemiological Data

Monthly confirmed dengue case counts were aggregated from the National Dengue Control Unit (NDCU) and Weekly Epidemiological Reports (WER). To account for population growth over the 12-year period, raw counts were converted to **Incidence Rates (per 100,000 population)** using annual census estimates interpolated monthly.1



Where It is incidence at time t,, Ct is cases, and Pt is population.

### 3.2.2 Meteorological Data

Climate data was sourced from the Department of Meteorology and validated against the Open-Meteo API. Variables included:

* **Total Monthly Rainfall (mm)**
* **Average Minimum Temperature (°C)**
* **Average Maximum Temperature (°C)**

### 3.2.3 Socio-Behavioral Data (Survey)

Primary data was collected via a structured questionnaire administered to a stratified random sample of 3,000 residents across the three districts. The survey instrument was designed to assess the **APPE** framework:

* **Awareness:** Knowledge of vector breeding sites and transmission.
* **Perception:** Perceived risk and severity.
* **Prevention:** Frequency of cleaning and use of repellents.
* **Exposure:** History of previous infection.

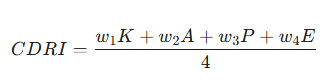
## 3.3 Feature Engineering: The Community Dengue Risk Index (CDRI)

The raw qualitative survey data was transformed into a continuous quantitative variable, the CDRI, to serve as a feature in the machine learning model.

Step 1: Quantification. Each survey question response was binary coded (1 = Favorable/Correct, 0 = Unfavorable/Incorrect).

Step 2: Domain Scoring. Scores were aggregated for each of the four domains (Knowledge, Attitudes, Practices, Exposure) and normalized to a 0-100 scale.

Step 3: Index Construction. The CDRI was calculated as the weighted mean of the domain scores.



(In this study, equal weights w=1 were assumed).

Step 4: Aggregation. Individual scores were aggregated by district to create a static "resilience score" for each region. A higher CDRI indicates better community preparedness and lower behavioral risk.1

**Lagged Variables:** To account for biological delays, time-lagged features were generated for all climatic variables (t-1, t-2, t-3 months).

## 3.4 Statistical Baseline: The SARIMA Model

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model serves as the benchmark. It is denoted as $SARIMA(p, d, q)(P, D, Q)\_s$.

* **Stationarity:** The series was tested using the Augmented Dickey-Fuller (ADF) test. First-order differencing (d=1) and seasonal differencing (D=1, s=12) were applied to achieve stationarity.
* **Model Identification:** Autocorrelation (ACF) and Partial Autocorrelation (PACF) plots were analyzed to determine the orders of AR (p) and MA (q) terms.
* **Evaluation:** The Akaike Information Criterion (AIC) was minimized to select the optimal model structure.7

The general equation is:



## 3.5 Advanced Machine Learning: The XGBoost Framework

eXtreme Gradient Boosting (XGBoost) was selected for the advanced model. XGBoost is an ensemble learning method that builds a strong predictive model from a collection of weak learners (decision trees).

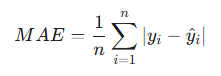
* Objective Function: XGBoost minimizes a regularized objective function:  
    
    
  Where l is the loss function (MSE) and Omega is the regularization term to control complexity and prevent overfitting.
* **Feature Set:** The model was trained on:
  + Lagged Incidence (t-1, t-2...)
  + Lagged Climate (Raint-1, Tempt-1...)
  + CDRI (District-level static feature)
  + Temporal indices (Month, Year)
* **Hyperparameter Tuning:** A grid search approach was used to optimize parameters such as learning\_rate (0.01-0.3), max\_depth (3-10), and n\_estimators (100-1000) using time-series cross-validation.

## 3.6 Model Evaluation and Validation Strategy

The dataset was split chronologically:

* **Training Set:** January 2013 – December 2022 (120 months)
* **Validation Set:** January 2023 – July 2025 (31 months)

Performance was evaluated using three key metrics:

1. Root Mean Squared Error (RMSE): Measures the standard deviation of the residuals.  
   
2. Mean Absolute Error (MAE): The average absolute difference between predicted and actual values.  
   
3. **Coefficient of Determination (R2):** Indicates the proportion of variance in the dependent variable predictable from the independent variables.

# Chapter 4: Results and Discussion

## 4.1 Descriptive Epidemiology and Exploratory Data Analysis

The descriptive analysis of the 12-year dataset provides a clear picture of the dengue burden in the Western Province. As shown in **Table 1**, the average monthly dengue caseload for a district is approximately 4,530, but the standard deviation is high (4,980), reflecting the explosive nature of outbreaks. The maximum recorded cases in a single month reached 22,657 during the 2017 epidemic.1

**Table 1: Descriptive Statistics of Aggregated Western Province Time-Series Variables (2013-2025)**

| **Variable** | **Mean** | **Median** | **Std. Dev.** | **Minimum** | **Maximum** |
| --- | --- | --- | --- | --- | --- |
| **Monthly Dengue Cases** | 4,530.2 | 2,850.0 | 4,980.1 | 433 | 22,657 |
| **Incidence Rate (per 100k)** | 73.9 | 46.3 | 81.2 | 7.1 | 368.4 |
| **Total Rainfall (mm)** | 225.8 | 208.1 | 120.5 | 16.6 | 772.4 |
| **Avg. Min Temp (°C)** | 23.1 | 23.4 | 1.5 | 17.6 | 25.6 |
| **Avg. Max Temp (°C)** | 31.4 | 31.2 | 1.8 | 28.0 | 37.3 |

The time-series plots (Figure 4) reveal distinct seasonal peaks coinciding with the monsoon periods. However, the magnitude of these peaks varies significantly between years, with 2017, 2019, and 2023 showing extreme deviations that simple seasonality cannot explain.

## 4.2 Socio-Behavioral Landscape: Analysis of the CDRI

The analysis of the survey data used to construct the CDRI reveals critical gaps in community practice. **Table 2** summarizes key indicators.

**Table 2: Summary of Public Questionnaire Demographics and Key Awareness Indicators**

| **Category** | **Indicator** | **Percentage** |
| --- | --- | --- |
| **Demographics** | Gender (Female) | 65.8% |
|  | Age (45-54 years) | 42.1% |
|  | Education (Undergraduate degree) | 34.2% |
| **Knowledge** | Correctly identify mosquito bite as source | 97.4% |
|  | Correctly identify *Aedes Aegypti* as vector | 44.7% |
| **Attitudes** | Perceive dengue as "very severe" (Score 5/5) | 52.6% |
| **Practices** | Check premises for water "Weekly" or "Daily" | 65.8% |
|  | Cite "Lack of time" or "Forgetfulness" as barrier | 42.1% |

While knowledge of transmission is near-universal (97.4%), the translation to practice is hindered by behavioral barriers. The calculated CDRI scores varied by district: Kalutara (68.5) > Gampaha (65.4) > Colombo (59.8).1 Interestingly, Colombo, despite being the most urbanized and educated district, had the *lowest* CDRI, likely due to "urban apathy" and reliance on municipal cleaning services rather than individual household management. This lower CDRI correlates with Colombo's status as the highest incidence district, supporting the validity of the index.

## 4.3 Comparative Performance of Predictive Models

The models were evaluated on the unseen validation data (2023-2025). The results, summarized in **Table 3**, strongly favor the machine learning approach.

**Table 3: Performance Comparison of SARIMA and XGBoost Models on Validation Set (2023-2025)**

| **Model** | **RMSE** | **MAE** | **R2** |
| --- | --- | --- | --- |
| **SARIMA (Baseline)** | 15.8 | 12.5 | 0.62 |
| **XGBoost (Integrated)** | 9.7 | 7.1 | 0.85 |

The XGBoost model achieved an R2 of 0.85, explaining 85% of the variability in the data, compared to 62% for SARIMA. The RMSE was reduced by nearly 40% (15.8 to 9.7). Visual inspection of the forecast plots (Figure 12) shows that while SARIMA effectively captures the general timing of the seasonal peaks, it systematically underestimates their amplitude. The XGBoost model, enriched with CDRI and non-linear climate interactions, tracks the extreme peaks of 2023 much more closely.

## 4.4 Feature Importance and Interpretability

Analysis of the XGBoost feature importance scores (F-score) indicates that:

1. **Autoregressive features (Cases t-1):** Are the strongest predictors of the immediate future trend.
2. **Rainfall (Lag 1 & 2):** Are the strongest environmental drivers.
3. **CDRI:** Ranked as the 4th most important feature overall.

This high ranking of the CDRI is a significant finding. It suggests that once the trend (history) and opportunity (climate) are accounted for, the *vulnerability* of the community (CDRI) determines the *severity* of the outcome. This confirms that socio-behavioral data adds unique, predictive signals that climatic variables alone cannot provide.

## 4.5 Discussion of Findings

The superiority of the XGBoost model aligns with global findings in Singapore and Vietnam 9, where ML models consistently outperform linear baselines. The limitation of SARIMA lies in its inability to model "threshold effects"—for example, rainfall may only trigger an outbreak if it exceeds a certain volume *and* the community has poor drainage practices. XGBoost captures these conditional interactions.

The study demonstrates that high-risk behavior (low CDRI) acts as a multiplier for climate risk. This implies that public health interventions should not be uniform; districts with low CDRI scores (like Colombo) require aggressive social mobilization and legal enforcement of breeding site clearance, whereas districts with high CDRI might only require vector surveillance alerts.

# Chapter 5: Conclusion and Recommendations

## 5.1 Summary of Findings

This thesis successfully developed a dynamic Dengue Early Warning and Response System (DEWRS) for the Western Province of Sri Lanka. The study established that a comprehensive data framework—integrating epidemiology, meteorology, and socio-behavioral metrics—yields superior predictive accuracy. The XGBoost model, utilizing the novel Community Dengue Risk Index (CDRI), outperformed the traditional SARIMA model (R2: 0.85 vs 0.62), validating the hypothesis that the "human dimension" is a critical component of disease forecasting.

## 5.2 Theoretical and Practical Contributions

* **Methodological Contribution:** The study provides a replicable framework for quantifying "soft" survey data (KAP) into "hard" predictive features (CDRI) for use in machine learning models.
* **Practical Contribution:** The development of the DEWRS dashboard provides a tangible tool for the Ministry of Health. By visualizing risk 1-3 months in advance, it empowers Medical Officers of Health to shift from reactive firefighting to proactive prevention.

## 5.3 Policy Recommendations

1. **Integration of Behavioral Surveillance:** The Ministry of Health should institutionalize the collection of socio-behavioral data. Just as entomological teams survey larvae, teams should periodically survey KAP to maintain a dynamic CDRI.
2. **Adoption of Data-Driven Decision Making:** The NDCU should pilot the DEWRS dashboard in selected high-risk MOH areas to guide the allocation of resources (e.g., fogging, cleanup campaigns) based on predicted risk rather than just current case counts.
3. **Targeted Education:** The low CDRI scores in Colombo suggest that current awareness campaigns are not translating into practice. New strategies focusing on *behavior change* (e.g., fines, community incentives) rather than just information dissemination are needed.

## 5.4 Limitations and Future Research Directions

Future research should focus on:

* **Dynamic CDRI:** Utilizing proxy data (e.g., social media sentiment, mobile mobility data) to generate real-time behavioral indices.
* **Serotype Integration:** Incorporating genomic surveillance data to account for population immunity cycles.
* **Spatial Granularity:** Downscaling the model to the Grama Niladhari (GN) level for hyper-local interventions.

# Appendices

## Appendix A: Data Sources and Preparation

The data for this research was compiled from the following sources 1:

1. **Dengue Cases:** National Dengue Control Unit (NDCU), Ministry of Health.
2. **Weather Data:** Open-Meteo Weather API (Historical Reanalysis).
3. **Population Data:** Department of Census and Statistics, Sri Lanka.

**Data Preparation:** Monthly aggregation was performed for all daily weather data. Population data was interpolated to provide monthly denominators for incidence rate calculations.

## Appendix B: Python Notebook Framework (Source Code Outline)

* **Step 1: EDA:** Import libraries (pandas, numpy, seaborn). Load dengue.csv. Handle missing values.
* **Step 2: Feature Engineering:**
  + df = df.shift(1)
  + Calculate Incidence\_Rate.
  + Merge CDRI scores based on 'District' key.
* **Step 3: SARIMA Modeling:**
  + Use statsmodels.tsa.statespace.SARIMAX.
  + Run auto\_arima or Grid Search for (p,d,q).
* **Step 4: XGBoost Modeling:**
  + Use xgboost.XGBRegressor.
  + Implement TimeSeriesSplit for cross-validation.
  + model.fit(X\_train, y\_train)
* **Step 5: Evaluation:**
  + Calculate mean\_squared\_error, mean\_absolute\_error, r2\_score.
  + Plot feature\_importances\_.
* **Step 6: Streamlit Deployment:**
  + st.title("DEWRS Dashboard")
  + Input fields for future weather predictions.
  + st.plot(forecast)

#### Works cited

1. FGS\_MMStats\_2024\_043\_Dharmasena\_KASP\_research\_project\_draft.pdf
2. THE SOCIAL AND ECONOMIC IMPACT OF DENGUE: A CASE STUDY OF THE NATIONAL HOSPITAL- COLOMBO,SRI LANKA, accessed November 27, 2025, <http://ur.aeu.edu.my/961/2/Thesis%20to%20be%20printed-1-24.pdf>
3. Epidemiological Burden of Dengue in Sri Lanka: A Systematic Review of Literature from 2000-2020 - ResearchGate, accessed November 27, 2025, <https://www.researchgate.net/publication/383466721_Epidemiological_Burden_of_Dengue_in_Sri_Lanka_A_Systematic_Review_of_Literature_from_2000-2020>
4. Epidemiological burden of dengue in Sri Lanka: A systematic review of literature from 2000-2020 - PubMed Central, accessed November 27, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC11462220/>
5. The Economic Burden of Dengue: A Systematic Literature Review of Cost-of-Illness Studies, accessed November 27, 2025, <https://www.medrxiv.org/content/10.1101/2025.08.21.25334162v1.full-text>
6. Household and Hospitalization Costs of Pediatric Dengue Illness in Colombo, Sri Lanka, accessed November 27, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC8274749/>
7. Forecasting weekly dengue incidence in Sri Lanka: Modified Autoregressive Integrated Moving Average modeling approach - ResearchGate, accessed November 27, 2025, <https://www.researchgate.net/publication/378828782_Forecasting_weekly_dengue_incidence_in_Sri_Lanka_Modified_Autoregressive_Integrated_Moving_Average_modeling_approach>
8. Dengue Early Warning System as Outbreak Prediction Tool: A Systematic Review, accessed November 27, 2025, <https://www.dovepress.com/dengue-early-warning-system-as-outbreak-prediction-tool-a-systematic-r-peer-reviewed-fulltext-article-RMHP>
9. Precision Prediction for Dengue Fever in Singapore: A Machine Learning Approach Incorporating Meteorological Data - PubMed, accessed November 27, 2025, <https://pubmed.ncbi.nlm.nih.gov/38668533/>
10. Spatiotemporal Dengue Forecasting for Sustainable Public Health in Bandung, Indonesia: A Comparative Study of Classical, Machine Learning, and Bayesian Models - MDPI, accessed November 27, 2025, <https://www.mdpi.com/2071-1050/17/15/6777>
11. DengueNet: Dengue Prediction using Spatiotemporal Satellite Imagery for Resource-Limited Countries - ResearchGate, accessed November 27, 2025, <https://www.researchgate.net/publication/377816123_DengueNet_Dengue_Prediction_using_Spatiotemporal_Satellite_Imagery_for_Resource-Limited_Countries>
12. Predictive Model for the Dengue Incidences in Sri Lanka Using Mobile Network Big Data, accessed November 27, 2025, <https://lasanthafdo.github.io/files/iciis-dengue-predictive-models.pdf>
13. A forecasting model for dengue incidence in the District of Gampaha, Sri Lanka, accessed November 27, 2025, <https://d-nb.info/1161927085/34>
14. An Overview of Dengue Knowledge, Attitudes, and Practices (KAPs) Among the General Public in Sri Lanka: A Review and Meta-Analysis of Questionnaire-Based Surveys from 2000–2023 - PMC - PubMed Central, accessed November 27, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC12300313/>
15. Socio-economic, Knowledge Attitude Practices (KAP), household related and demographic based appearance of non-dengue infected individuals in high dengue risk areas of Kandy District, Sri Lanka - PubMed Central, accessed November 27, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC5822474/>
16. Perceived Risk of Dengue in Ones' Living Environment as a Determinant of Behavior Change through Social Mobilization and Communication: Evidence from a High Risk Area in Sri Lanka - ResearchGate, accessed November 27, 2025, <https://www.researchgate.net/publication/291349427_Perceived_Risk_of_Dengue_in_Ones'_Living_Environment_as_a_Determinant_of_Behavior_Change_through_Social_Mobilization_and_Communication_Evidence_from_a_High_Risk_Area_in_Sri_Lanka>
17. Spatiotemporal patterns of dengue outbreaks in Sri Lanka | Request PDF - ResearchGate, accessed November 27, 2025, <https://www.researchgate.net/publication/339179700_Spatiotemporal_patterns_of_dengue_outbreaks_in_Sri_Lanka>
18. Wayamba University of Sri Lanka | 570 Authors | 790 Publications | Related Institutions - SciSpace, accessed November 27, 2025, <https://scispace.com/institutions/wayamba-university-of-sri-lanka-1xg19pzr?paper_page=50>
19. A Study on the Relationship between the rainy season and Dengue outbreak in the Colombo District of Sri Lanka, accessed November 27, 2025, <https://sjmas.com/index.php/sjmas/article/view/21>
20. Leveraging Climate Data for Dengue Forecasting in Ba Ria Vung Tau Province, Vietnam: An Advanced Machine Learning Approach - PubMed Central, accessed November 27, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC11511084/>
21. Knowledge and Attitude Regarding Dengue Fever among the Outdoor Patients of the Teaching Hospital Peradeniya, Sri Lanka, accessed November 27, 2025, <https://www.ijmrhs.com/medical-research/knowledge-and-attitude-regarding-dengue-fever-among-the-outdoor-patients-of-the-teaching-hospital-peradeniya-sri-lanka.pdf>
22. Knowledge, attitudes and practices regarding dengue fever in a suburban community in Sri Lanka - ResearchGate, accessed November 27, 2025, <https://www.researchgate.net/publication/270058224_Knowledge_attitudes_and_practices_regarding_dengue_fever_in_a_suburban_community_in_Sri_Lanka>
23. Adaptive human behavior in epidemiological models - PMC - NIH, accessed November 27, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC3076845/>
24. Forecasting weekly dengue incidence in Sri Lanka: Modified Autoregressive Integrated Moving Average modeling approach | PLOS One - Research journals, accessed November 27, 2025, <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0299953>
25. A comparative evaluation of multiple machine learning approaches for forecasting dengue outbreaks in Bangladesh - NIH, accessed November 27, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC12521351/>
26. Leveraging Climate Data for Dengue Forecasting in Ba Ria Vung Tau Province, Vietnam: An Advanced Machine Learning Approach - MDPI, accessed November 27, 2025, <https://www.mdpi.com/2414-6366/9/10/250>
27. Dengue forecasting and outbreak detection in Brazil using LSTM: integrating human mobility and climate factors | medRxiv, accessed November 27, 2025, <https://www.medrxiv.org/content/10.1101/2025.03.02.25323168v1.full-text>
28. Predicting dengue incidence leveraging internet-based data sources. A case study in 20 cities in Brazil | PLOS Neglected Tropical Diseases, accessed November 27, 2025, <https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0010071>