MJSAT

Malaysian Journal of Science and Advanced Technology

journal homepage: https://mjsat.com.my/

Enhancing Dengue Outbreak Prediction in Bangladesh: A Weighted Average Ensemble Machine Learning Approach

Mahfujur Rahman*[1], Noboranjan Dey[1], Nazmus Sakib[1], Rukaiya Jahan Sajuti[2], Mehedi Hasan[3], Abdullah Hel Azmain[1], Rahul Biswas[1], Dipta Gomes[1], Kazi Tanvir[1], and Mirza Asif Mahmud[1]

[1]*Department of Computer Science, American International University-Bangladesh (AIUB), Dhaka, Bangladesh.*
[2]*Department of Management Studies, Jagannath University (JnU), Dhaka, Bangladesh.*
[3]*Department of Electrical & Electronic Engineering, American International University-Bangladesh (AIUB), Dhaka, Bangladesh.*

KEYWORDS

Dengue Outbreak Prediction
Ensemble Learning
Machine Learning Models
Public Health Informatics
Dengue Transmission

ARTICLE HISTORY

Received 3 September 2025
Received in revised form 6 May 2026
Accepted 15 June 2026
Available online 24 June 2026

ABSTRACT

Dengue fever is one of the most urgent public health threats in Bangladesh, particularly in cities like Chattogram, where rapid urban growth, poor waste management, and changing climate conditions create a fertile ground for outbreaks. However, most existing dengue prediction models lack contextual adaptation, often relying on single classifiers that fail to capture localized socio-environmental factors, limiting their predictive reliability in the Bangladeshi context. To strengthen early prediction and response, this study introduces a Weighted Average Ensemble Learning (WAEL) model that combines the strengths of Support Vector Machine (SVM), Random Forest (RF), and AdaBoost classifiers. Using a dataset of 199 records and nine attributes collected from a Figshare survey on dengue awareness and prevention, extensive preprocessing steps such as feature selection, cleaning, and imputation were applied to ensure high data quality. Six baseline machine learning models, including Decision Tree, Naïve Bayes, k-Nearest Neighbor, SVM, RF, and AdaBoost, were evaluated, with RF and AdaBoost emerging as the strongest individual performers, each achieving 92.5% accuracy and F1-scores above 0.95. The proposed WAEL approach surpassed all individual models, achieving 93% accuracy, 94% precision, 98.5% recall, and an F1-score of 0.955. These findings demonstrate the advantage of ensemble methods in producing more reliable and context-aware predictions by harnessing the complementary strengths of multiple classifiers. Beyond technical performance, the study offers valuable insights for policymakers and health authorities to identify high-risk areas, vulnerable populations, and key factors driving dengue transmission, ultimately providing a data-driven framework for targeted prevention and early intervention in Bangladesh.

1. INTRODUCTION

Dengue fever (DF), a mosquito-borne viral disease caused by the Dengue Virus (DENV), poses a significant public health challenge worldwide, particularly in tropical and subtropical regions. Over 3.9 billion people globally are at risk of dengue infection, with the World Health Organization (WHO) classifying dengue as one of the most critical emerging infectious diseases [1]. The disease is primarily transmitted by Aedes mosquitoes, which thrive in urban areas characterized by high population density, poor waste management, and favourable climatic conditions, such as high temperatures, rainfall, and humidity. These factors create ideal breeding environments for mosquitoes, contributing to the cyclical and complex transmission dynamics of the disease.

DF has become a growing public health concern in Bangladesh, with urban areas such as Chattogram experiencing severe outbreaks. Chattogram, the country's second-largest city, is particularly vulnerable because of rapid urbanization, inadequate waste management, and climate

*Corresponding author:
E-mail address: Mahfujur Rahman< mahfuj@aiub.edu >.
https://doi.org/10.56532/mjsat.v6i2.601

variability. The 2019 outbreak was the worst on record for Bangladesh, with over 100,000 reported cases and 164 confirmed deaths [2]. Chattogram contributes significantly to this national burden, with sharp increases in hospital admissions. In subsequent years, the city continued to experience troubling trends in its economy.

According to data from the Civil Surgeon's Office, the incidence of dengue cases in Chattogram has increased significantly in recent years. In 2020, only 17 cases were reported [3], but this number surged to 271 in 2021, resulting in five fatalities [4]. The situation deteriorated further in 2022, with a sharp rise in 5,445 reported cases and 41 deaths [5]. By 2023, 239 individuals had been diagnosed with dengue, with three fatalities recorded [6]. Data from the Directorate General of Health Services (DGHS) for 2023 indicate that the Dhaka district experienced the highest dengue outbreak, with 113,233 confirmed cases and 981 deaths, marking the highest mortality rate [7]. That same year, the Chattogram district was the second most severely affected region, with more than 14,000 reported infections and over 100 fatalities [8]. These alarming statistics emphasize the critical need for proactive measures to control the dengue epidemic in Chattogram. Targeted preventive interventions are crucial in high-risk areas such as Chattogram, particularly as climate change and rapid urbanization continue to exacerbate dengue spread.

The clinical presentation of DF is diverse, ranging from mild febrile illness to severe forms, such as dengue haemorrhagic fever (DHF) and dengue shock syndrome (DSS), which can lead to significant morbidity and mortality if not managed promptly. Certain demographic groups, including individuals aged 60 years, men, workers, homemakers, and students, are particularly vulnerable. Additionally, behavioural factors, such as smoking and alcohol consumption, may contribute to this risk. However, these observations are often based on limited sample sizes, underscoring the need for further research to confirm these findings and establish causal relationships between risk factors and disease.

Our study aimed to comprehensively analyse the risk factors, transmission patterns, and seasonal trends of dengue to address these challenges. In summary, the contributions of this study are as follows:

- Developed a clean and well-structured dataset using systematic preprocessing, including feature selection, cleaning, and imputation, to ensure model reliability.
- Evaluated six machine learning models (DT, NB, kNN, SVM, RF, and AdaBoost) to identify the most effective predictors of dengue risk and awareness.
- Proposed a Weighted Average Ensemble Learning (WAEL) model that combines SVM, RF, and AdaBoost, achieving superior accuracy and robustness compared to individual classifiers.
- Provided actionable insights for public health authorities to support targeted dengue prevention and awareness strategies in high-risk areas.

The remainder of this paper is structured as follows: Section 2 reviews the related literature, highlighting previous studies on dengue risk factors, transmission patterns and predictive modelling approaches. Section 3 describes the research methodology, including data collection, preprocessing, feature selection, and ML models used in the study. Section 4 discusses the experimental setup and presents the results of the model evaluation. Finally, Section 5 concludes the study with key insights, policy implications and potential directions for future research.

2. LITERATURE REVIEW

DF remains a significant global public health challenge, with its transmission dynamics influenced by climatic, environmental, and sociodemographic factors. Recent advancements in ML and statistical modeling have enhanced our ability to predict outbreaks, identify high-risk areas and evaluate intervention strategies. This study provides a comprehensive analysis of the existing literature on ML and computational techniques applied to dengue transmission and prevention, focusing on methodologies, key contributions, datasets, advantages, and limitations. A summary of previous research studies and their significant findings is presented in Table 1.

Table 1. Summary of literature review for identifying dengue disease transmission and prevention measures

Studies	Methods	Key Contributions	Dataset Information	Significance	Limitations
Cortes, C. and Vapnik, V. [9]	SVNs, quadratic programming, kernel functions	Binary classification, high generalization	Synthetic 2D data, USPS & NIST databases	Outperforms traditional methods, handles high dimensions	Computationally intensive, kernel-dependent
Chanprasopchai, P. et al. [10]	SIR model with vaccination effects	Dengue transmission dynamics, vaccination impact	Epidemiological data (available on request)	Enhances control efforts, public health foundation	Simplified assumptions lack real-world complexity
Nayak, M.S.D.P. and Narayan, KA [11]	Seasonal ARIMA (1, 0, 0) $(0, 1, 1)_{12}$	Monthly dengue forecasting in Kerala	Secondary data (2006–2018, PDF reports)	Strong predictive performance	Relies on secondary data, ignores external factors
Kakarla, S.G. et al. [12]	DLNM for climatic effects	Seasonal dengue patterns (August–September peak)	Climate & Epidemiological Data (2010–2017)	Links climate to transmission	Omits socio-demographic factors, limited spatial granularity
Mala, S. and Jat, M.K.	Space-time scan	Spatio-temporal clusters	Daily cases (2010–	High-risk cluster	The short study period excludes

[13]	statistics, GIS techniques	in Delhi, wind-speed association	2012), wind speed data	detection, resource allocation	socio-economic factors
Ho, T-S et al. [14]	DNN, LR, DT	Clinical prediction using minimal variables (age, platelets)	Lab-confirmed cases (2015, Taiwan)	Reliable outcomes with few inputs	Excludes co-infections (e.g., malaria)
Munarsih, E. and Saluza, I. [15]	ARIMA (2,1,2) vs. Exponential Smoothing	ARIMA outperforms (lower MSE/MAE)	Annual cases (2003–2016, Palembang)	Superior accuracy aids public health planning	Manual tuning ignores climate/demographics
Hoyos, W. et al. [16]	RF, GWR, ANN, SVM	Diagnostic, epidemic, and intervention modeling	Multi-source (clinical, climate, social media)	Comprehensive ML overview, federated learning potential	Data quality issues, language bias
Yusuff, M. [17]	GIS hotspot analysis (KDE, Getis-Ord Gi*)	Dengue-prone areas, multi-source integration	Health, satellite, and meteorological data	Early risk detection supports decision-making	Ethical/technical barriers, real-time challenges
Sarker, I. and Karim, M.R. [18]	GAMLSS with B-Spline	Climatic predictors (humidity, temperature)	Daily cases, climate data (2020–2021, Dhaka)	Data-driven prevention strategies	Linear precipitation assumption, 2-year data limit
Ramírez-Soto, M.C. et al. [19]	SIR-SI with temperature-dependence	Climate-integrated outbreak prediction	Weekly cases (2016–2020, Peru), NOAA data	Improve accuracy in low-transmission areas	Small sample size, homogeneous assumptions
Clarke, J. et al. [20]	OpenDengue database	Globally standardized dengue data (102 countries)	Multi-source (1924–2023, 56M cases)	High-resolution, public access	Reporting biases, manual processing needs

Several studies have employed advanced ML algorithms to classify dengue cases and predict outbreaks. Cortes, C. and Vapnik, V. [9] have applied Support Vector Networks (SVNs) as a binary classification technique to demonstrate high generalization ability in high-dimensional spaces. However, their computational complexity and dependence on kernel functions limit their scalability. Ho et al. [14] used Deep Neural Networks (DNNs), Logistic Regression (LR), and DTs models on clinical data. In their study, they achieved reliable prediction results using minimal input variables (e.g., age and platelet count). A notable limitation of this study was the exclusion of co-infections (e.g., malaria), which may have confounded the results. Another research performed by Hoyos et al. [16] applied RF and Geographically Weighted Regression (GWR) methods in their experiments. They provide robust diagnostic and epidemic modeling, although data quality and language biases in sourcing studies are challenges.

Another primary technique is time-series analysis, which has been pivotal in predicting dengue outbreaks. Nayak, M.S.D.P. and Narayan, KA [11] applied an innovative time series analysis and forecasting method named Seasonal ARIMA. In this study, forecasting monthly dengue cases in Kerala, India, outperformed other ARIMA analyses but relied on secondary data, which may result in risky verification. Another study by Munarsih, E. and Saluza, I. [15] also applied the ARIMA method and showed superior results with high accuracy and low error matrices. However, manual parameter tuning and the exclusion of external factors (e.g., climate) reduce generalizability.

Geospatial techniques are essential for identifying high-risk zones and transmission patterns. Mala, S. and Jat, M.K. [13] detect spatio-temporal clusters in Delhi using Kulldorff's Space-Time Scan Statistic and GIS techniques, integrating demographic and wind-speed data. Their research used a short study period (3 years) and excluded socioeconomic factors, which may have created biases in their outcomes. Yusuff, M. [17] also applied GIS-Based Hotspot Analysis techniques by combining multi-source data (e.g., satellite imagery) for early risk detection but faced real-time integration challenges and ethical concerns.

Transmission dynamics and climatic effects play significant roles in minimizing dengue hazards. Chanprasopchai et al. [10] and Ram´ırez-Soto et al. [19] used SIR and SIR-SI models to evaluate the impact of vaccination and temperature-dependent transmission of this epidemic disease. However, they applied simplified assumptions (e.g., homogeneous interactions) and small sample sizes, which minimized the applicability of their studies. Kakarla, S.G. et al. [12] applied the Distributed Lag Non-Linear Models (DLNM) method in their research to quantify delayed climatic effects (e.g., rainfall) but omitted socio-demographic variables, which the other researchers declared is as a very critical feature for dengue transmission. Sarker, I. and Karim, M.R. [18] identified humidity and temperature in the Dhaka region as key predictors and applied GAMLSS with B-Spline Regression in their research. Their study used linear assumptions for precipitation effects, which may oversimplify the relationships.

Clarke has performed another research, J. et al. [20], to standardize global dengue data from 102 countries, enabling high-resolution analyses using the OpenDengue method. Although better public health decision-making outcomes have

been found in their research, variability in reporting standards and manual processing requirements pose challenges.

In every ML study, a standard dataset plays a vital role in analyzing and generating trustworthy and reliable outcomes. Clinical and Laboratory Data used by Ho, T-S et al. [14] and Ram´ırez-Soto, M.C. et al. [19] which included confirmed cases and climatic variables but often lacked granularity (e.g., serotype details). Kakarla, S.G. et al. [12] and Sarker, I. & Karim, M.R. [18] integrate epidemiological and environmental climate records but exclude urban-specific factors (e.g., sanitation). Although ethical and technical barriers hindered scalability, Yusuff, M. [17] combines multi-source geospatial health records with satellite imagery.

In summary, previous studies have significantly advanced dengue outbreak prediction using statistical, epidemiological, and machine learning approaches. However, several overarching challenges persist. Many existing models rely on single algorithms (e.g., ARIMA, SVM, DNN) that exhibit limited generalization and are sensitive to local variations in climatic and socio-demographic factors. Time-series and compartmental models often employ simplified assumptions, neglecting non-linear interactions among environmental and behavioural variables. Likewise, most machine learning approaches are dataset-specific and fail to integrate multiple model perspectives, leading to overfitting and reduced predictive reliability in diverse regional contexts such as Bangladesh.

To address these limitations, the proposed Weighted Average Ensemble Learning (WAEL) model integrates the complementary strengths of SVM, Random Forest, and AdaBoost classifiers to enhance robustness, reduce bias, and improve predictive stability across varying input conditions. Unlike single-model frameworks, WAEL leverages model diversity and weighted aggregation to capture both linear and non-linear dengue transmission dynamics, offering a scalable and context-aware predictive tool tailored to Bangladesh's urban environments.

3. RESEARCH METHODOLOGY

DENV is a mosquito-borne disease that represents a significant global public health concern because of its extensive economic, social, and health-related impacts. This study employed a quantitative approach to examine DF patterns, risk factors, and transmission dynamics in Chattogram, Bangladesh. The methodology incorporates statistical and ML techniques to identify temporal trends, high-risk geographic areas and vulnerable demographic groups. By utilizing data-driven modeling, this study aims to improve the prediction of dengue outbreaks and support the development of effective public health interventions.

3.1 Dataset Overview

The dataset utilized in this study was sourced from research conducted by Rahman, Sahidur, et al., which focused on dengue vectors among dengue patients and the general population in Chattogram, Bangladesh [21]. The data were retrieved from the Figshare data repository titled "Knowledge and prevention practice regarding dengue fever" and contained comprehensive information regarding demographic characteristics, knowledge and awareness, prevention practices, environmental factors, and health outcomes. A summary of the dataset is presented in Table 2.

Table 2. Dengue virus awareness and prevention techniques dataset summary

Dataset Name	Data Source	No. of Attributes	No. of Instances	Type of Data
Knowledge and prevention practice regarding dengue fever [21]	Figshare	41	300	Tabular, Multivariate (Binary, Nominal, and Continuous)

The dataset comprises primary data collected from dengue patients and members of the general population to assess their knowledge and practices regarding the prevention of mosquito vectors responsible for transmitting DF.

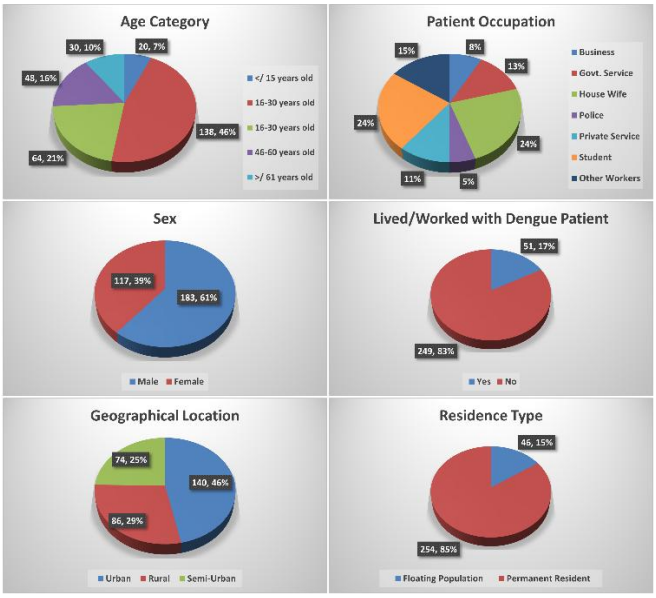

Fig. 1. The visual distribution of different categorical attributes of the dataset

The dataset included 300 entries and 41 variables, encompassing both categorical and numerical data. It is structured to facilitate the analysis of factors influencing dengue-related knowledge and preventive practices. The key demographic attributes captured in the dataset included age category, patient occupation, sex, contact history with dengue patients, geographical location, and type of residence. The target attribute of this dataset for the binary classification task is Total Knowledge Score based on responses measuring the knowledge level about dengue. Figure 1 presents a visual representation of the dataset and its characteristics in detail.

3.2. Data Preprocessing

Data preprocessing is a critical component of the research process that ensures the quality and suitability of the data for subsequent analysis. This phase encompasses several tasks, including data integration, selection, cleaning, transformation, feature selection, and engineering. These procedures are essential for converting raw data into structured and analyzable formats. Effective preprocessing enhances the interpretability and usability of the dataset, facilitates the extraction of relevant features, and supports the accurate

application of predictive modeling methods. Figure 2 illustrates the data preprocessing workflow and its implementation in this study.

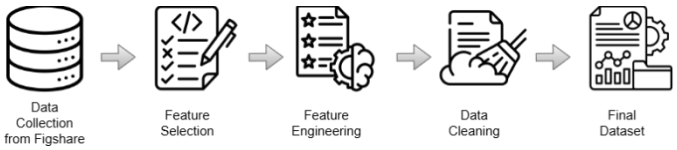

Fig. 2. Data preprocessing techniques

a) Feature Selection: Feature selection is a fundamental technique used to reduce the number of input variables by retaining only the most relevant data and eliminating noise. This involves identifying and isolating consistent, non-redundant, and significant features for model development. As datasets increase in size and complexity, systematic feature selection becomes more important. The primary objectives of feature selection in this study were to simplify model structures, enhance performance, reduce computational costs, and mitigate the risk of overfitting, ultimately resulting in more accurate and efficient predictive models.

The dataset collected in this study comprised 300 instances, each with 41 attributes. Initially, 32 attributes were removed, including patient identification details, less informative variables, and derived attributes that introduced unnecessary complexity for ML applications. After this refinement, the dataset was reduced to nine attributes across 300 instances, forming the basis for the subsequent analysis and model development phases.

b) Feature Importance: To better understand the contribution of each variable in predicting dengue awareness and prevention knowledge, feature importance values were analyzed from the Random Forest model. The analysis indicated that contact history with dengue patients, type of residence, and age category were the most influential features, suggesting that individuals who had prior exposure to dengue cases, lived in densely populated or low-hygiene areas, and belonged to specific age groups exhibited more distinct awareness patterns. These variables demonstrated the strongest discriminatory power in determining the Total Knowledge Score.

Moderate influence was observed for occupation, sex, and educational level, reflecting their secondary but relevant impact on dengue-related knowledge and preventive practices. In contrast, features related to environmental cleanliness and household characteristics contributed marginally, likely due to homogeneity within the sampled population.

This finding underscores the interpretability of the model by linking demographic and behavioral factors with knowledge levels, reinforcing that the ensemble framework not only achieves high predictive accuracy but also yields contextually meaningful insights for targeted health awareness programs.

c) Feature Engineering: Feature engineering involved transforming raw data into a structured format to enhance model accuracy and reliability. A key focus was handling missing values to preserve data integrity and prevent biased predictions. The data set, collected from a dengue-related survey, initially contained 300 missing values. Two records with multiple missing attributes were removed, and 47 missing values were imputed using the Multiple Imputation by Chained Equations (MICE) method.

Prior to imputation, statistical inspection and pairwise correlation analysis were conducted to examine the missingness pattern. The results indicated that the missing values were dependent on variables such as occupation and residential location, classifying the mechanism as Missing at Random (MAR). This justified the use of MICE, as it effectively models conditional relationships among variables to produce unbiased estimates while maintaining the statistical integrity and distributional properties of the dataset. After imputation, the final dataset comprised 298 instances with nine attributes for subsequent model development.

d) Data Cleaning: Data cleaning is a foundational and critical step in ML research and a prerequisite for reliable data analysis. It involves the identification, correction, or removal of corrupted, duplicate, incomplete, inaccurate, or noisy records to enhance the overall quality of the dataset. The primary objective of data cleaning is to improve data accuracy and consistency, reduce bias, and ensure that each data point contributes uniquely and meaningfully to the learning process. This step significantly enhances the performance and reliability of ML models by eliminating redundancies and correcting inconsistencies.

In this study, data-cleaning procedures revealed 99 duplicate instances within the dataset. These duplicates were removed to ensure the integrity and dependability of the data for subsequent modeling and analyses. Following this process, the refined dataset consisted of 199 unique instances, which were used for in-depth analysis and model development.

After successfully completing data preprocessing, the final dataset comprised 199 instances and nine attributes. This refined dataset was free of missing values, redundant attributes, and duplicate entries, thereby ensuring its quality and suitability for analysis. The binary classification for the target variable was the Total Knowledge Score derived from the survey responses. These comprehensive preprocessing steps produced a clean, well-structured dataset optimized for accurate model training and meaningful analytical outcomes. The dataset is now fully prepared for implementing ML models and conducting further analysis.

e) Dataset Splitting: Dataset splitting is a critical step in machine learning that partitions data into training, validation, and testing subsets to support model training, hyperparameter tuning, and performance evaluation. This process helps prevent overfitting and ensures the model's ability to generalize to unseen data. The primary objective of dataset splitting is to promote model generalization and improve predictive reliability.

The dataset used in this study contained 199 instances with nine attributes after preprocessing. To optimize model performance, the data were divided using a structured approach. A total of 80% of the dataset, equivalent to 159 instances, was allocated for model development, while the remaining 20%, consisting of 40 instances, was reserved for testing. Within the 80% development subset, 75% of the data, approximately 119 instances or 60% of the total dataset, was used for training, and 25%, approximately 40 instances or 20% of the total dataset, was used for validation. This allocation ensured a balanced and systematic data distribution

with 60% for training, 20% for validation, and 20% for testing, providing a consistent framework for robust model development and evaluation. The dataset-splitting strategy adopted in this study is illustrated in Figure 3.

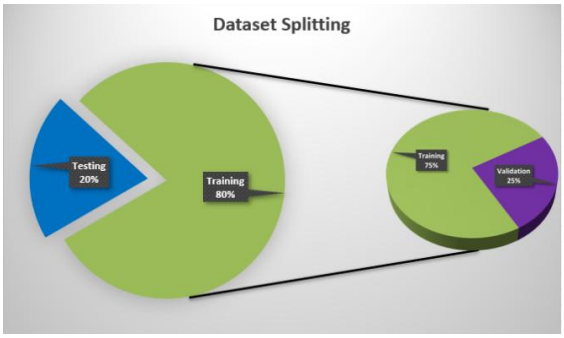

Fig. 3. The visualization of dataset splitting

3.3 Proposed Methodology

This study aimed to develop an accurate and robust predictive model to analyze patterns in dengue transmission, focusing on Total Knowledge Score. Figure 4 depicts the overall system architecture designed to analyze patterns in dengue transmission. It begins with the collection of raw data, which is gathered for analysis. This step involved refining the dataset by selecting important features, removing unnecessary attributes, addressing missing values, and eliminating duplicate data to enhance the model performance.

Once the data were pre-processed, widely used ML models were chosen, such as DT, SVM, kNN, AdaBoost, and RF, and their performances were tested for dengue pattern transmission. Following the evaluation, the three best-performing classification algorithms were selected based on the F-measure. The selected classifiers designed a novel EL classification approach using a weighted average mechanism. The following subsections provide a detailed explanation of the classification algorithms and the weighted average mechanisms.

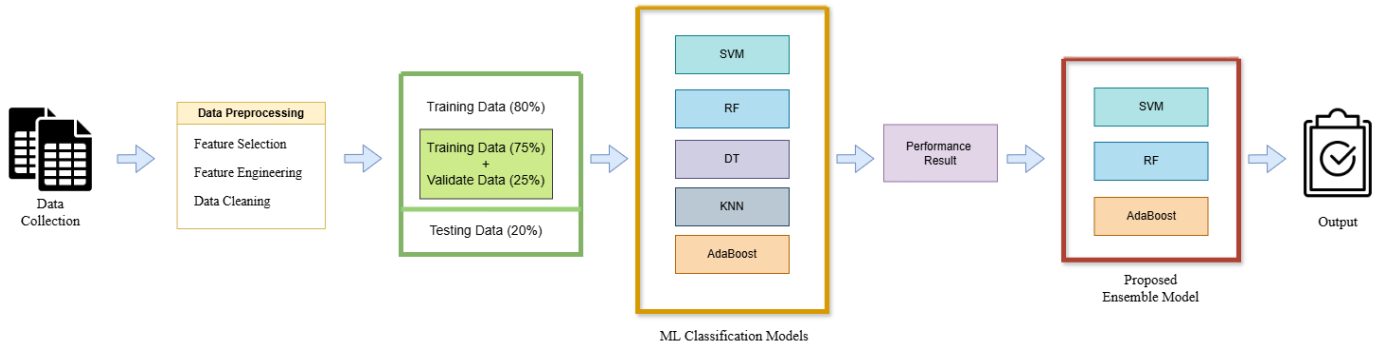

Fig. 4. Overall system architecture for analyzing dengue transmission patterns

- **Support Vector Machine (SVM):** SVM is a supervised learning algorithm used for classification tasks. It identifies the optimal hyperplane that separates data points into distinct classes while maximizing the margin between them [22]. This study implemented SVM to classify the Total Knowledge Score as a target attribute. The model achieved commendable accuracy and precision, thereby demonstrating its potential for predictive analysis.
- **k-Nearest Neighbor (kNN):** The kNN algorithm is a simple and effective non-parametric method used for classification tasks [23]. By predicting a data point's class based on the majority class of its k-nearest neighbors, kNN delivers high precision and recall in our analysis. The model performed well on the dataset, particularly with optimized 'k' values.
- **Random Forest (RF):** RF is a machine learning algorithm that aggregates the results of multiple DTs to produce a final output [24]. It is versatile and can be used for both classification and regression. In this study, RF consistently demonstrated strong predictive performance, achieving high recall scores, especially for the Total Knowledge Score. Its ability to handle categorical and continuous features, resist overfitting, and identify key attributes make it one of the most reliable standalone models.
- **Naïve Bayes (NB):** NB is a probabilistic classifier based on Bayes' theorem, which assumes feature independence [25]. In our study, it served as a baseline model, demonstrating moderate accuracy and providing valuable comparative insights. Despite its simplicity, the predictive capabilities of the algorithm were constrained by the interdependencies between the features in the dataset.
- **Decision Tree (DT):** DTs are intuitive models that split data based on feature values to make predictions [26]. This algorithm delivered high accuracy and interpretability, providing foundational insights into the key decision paths in the dataset. However, DTs tend to overfit, particularly with noisy or imbalanced data.
- **Adaptive Boosting (AdaBoost):** AdaBoost, short for Adaptive Boosting, combines multiple weak learners to form a strong predictive model [27]. In this study, AdaBoost performed admirably, achieving high accuracies and precisions. Its iterative weighting of challenging data points proved effective, although its sensitivity to outliers occasionally affected the results.
- **Weighted Average Ensemble Learning (WAEL) Technique:** The use of EL methods to combine multiple models has proven effective in enhancing prediction accuracy across various domains, such as classification and biometrics [28], [29]. EL is an ML approach that combines predictions from multiple base models to improve overall accuracy and robustness. The primary objective of EL methods is to achieve more accurate and dependable predictions than those of a single predictive model. In the proposed WAEL framework, the final

ensemble prediction was generated as a weighted combination of three base classifiers: SVM, RF, and AdaBoost. Figure 5 visually represents our proposed WAEL model, along with its technical details. The ensemble prediction is defined as:

$$\hat{y} = \sum_{i=1}^{n} w_i \cdot y_i \quad (1)$$

where $\hat{y}$ denotes the final ensemble output, w_i is the assigned weight for model i, and y_i represents the corresponding model prediction.

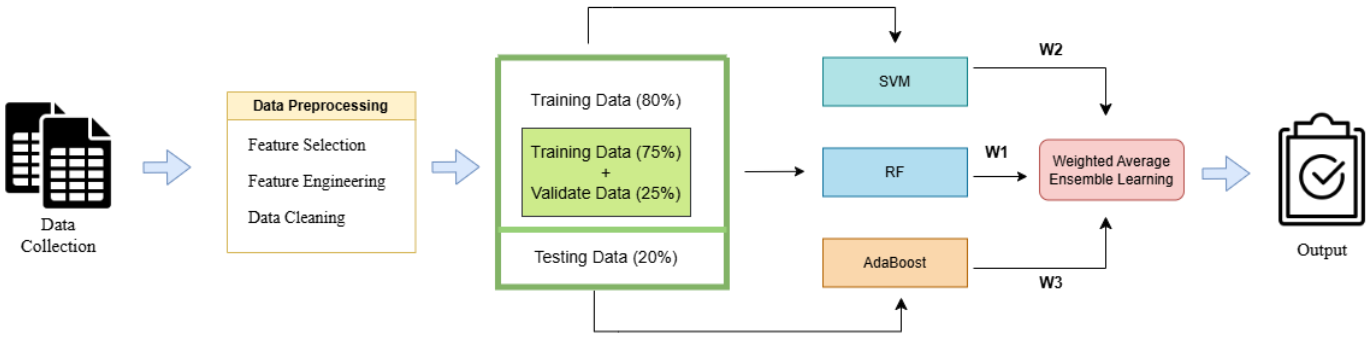

Fig. 5. Proposed Weighted Average Ensemble Learning (WAEL) model architecture

Weight Assignment Rationale:

The weights (w_i) were determined empirically using the validation phase of a 5-fold cross-validation. Each model's mean validation accuracy was normalized to compute proportional weights, ensuring that models with stronger generalization (higher validation accuracy) contributed more significantly to the final ensemble prediction. The sum of all weights was constrained to 1.0:

$$\sum_{i=1}^{n} w_i = 1 \quad (2)$$

The final optimized weights were as follows: RF = 0.337, SVM = 0.334, and AdaBoost = 0.329.

This weighted aggregation strategy effectively captures the complementary strengths of the individual classifiers while minimizing overfitting, resulting in a stable and interpretable ensemble framework.

Hyperparameter Comparison:

Table 3 summarizes the tuned hyperparameter ranges and the optimal values selected for each machine learning model, highlighting the parameter settings that yielded the best performance and model stability.

Table 3. Comparison between tuned and optimal hyperparameters

Model	Tuned Parameters (Search Range)	Optimal Values
SVM	Kernel type (linear, rbf); Penalty parameter C {0.1, 1, 10}; Kernel coefficient γ {0.01, 0.1, 1}	RBF kernel; C = 1; γ = 0.1
kNN	Number of neighbors k {3, 5, 7, 9, 11, 13, 15}; Distance metric (euclidean, manhattan)	k = 5; Euclidean distance
RF	Number of trees {50, 100, 200}; Maximum depth {5, 10, 15}; Criterion (gini, entropy)	100 estimators; Max depth = 10; Gini criterion
AdaBoost	Number of weak learners {50, 100, 200}; Learning rate {0.1, 0.5, 1.0}	100 weak learners; Learning rate = 0.5
DT	Max depth {3, 5, 10, 15}; Criterion (gini, entropy)	Max depth = 10; Gini criterion
NB	Smoothing parameter α {0.1, 0.5, 1.0}	α = 1.0

4. Experimental Results

This experimental evaluation aimed to assess how well different ML models could predict the total knowledge score from the regional survey responses in Chattogram. Each model was evaluated based on accuracy, precision, recall, and F1 score to provide a comprehensive understanding of its effectiveness. We aim to optimize the model performance through rigorous experimentation and identify the most effective approach for accurate classification.

4.1 Experimental Setup

All experiments in this study were conducted using Python 3.9 on the Scikit-learn (v1.2) machine learning framework, supported by Pandas, NumPy, and Matplotlib for data handling and visualization. Model training and evaluation were performed on a workstation with Intel Core i7 processor, 16 GB RAM, and Windows 11 environment.

4.2 Result Analysis

The performance analysis of various ML algorithms revealed key insights into their suitability for predicting an individual's knowledge of dengue from input data. This section elaborates on the results by comparing the strengths and limitations of each model to provide a comprehensive understanding of their effectiveness. Table 4 presents the performance metrics of various ML models used in this study, including accuracy, precision, recall, and F1 score.

Table 4. Model performance comparison between proposed work and state-of-the-art techniques

Model	Accuracy	Precision	Recall	F1 Score
AdaBoost	0.925	0.9677	0.9375	0.9518
SVM	0.925	0.9394	0.9688	0.9536
RF	0.925	0.9143	1.0000	0.9545
*k*NN	0.900	0.9667	0.9063	0.9362
DT	0.900	0.9375	0.9375	0.9375
NB	0.850	0.9333	0.8750	0.9036
Proposed WAEL Model*	**0.930**	**0.9400**	**0.9850**	**0.9550**

Among the standalone models, AdaBoost, SVM, and RF demonstrated the highest accuracy of 92.5%, with F1 scores ranging from 0.9518 to 0.9545. kNN and DT followed closely, achieving an accuracy of 90%, while NB had the lowest

performance with an accuracy of 85% and an F1 score of 0.9036. Although these standalone models showed competitive results, their performances remained relatively similar.

To verify that the observed performance differences were not due to random variation, we conducted a 5-fold cross-validation for each model and calculated 95 % confidence intervals (CIs) for both accuracy and F1-score. The confidence intervals for the top three models were as follows: RF (Accuracy = 92.5 ± 1.8 %, F1 = 0.954 ± 0.012), SVM (Accuracy = 92.5 ± 2.1 %, F1 = 0.953 ± 0.015), and AdaBoost (Accuracy = 92.5 ± 2.0 %, F1 = 0.952 ± 0.014). The proposed WAEL model achieved 93.0 ± 1.6 % accuracy and F1 = 0.955 ± 0.010, confirming that its improvement, although numerically modest, falls within a statistically consistent range across folds. This analysis indicates that WAEL provides more stable and generalizable performance compared with single classifiers, reducing variance without compromising bias.

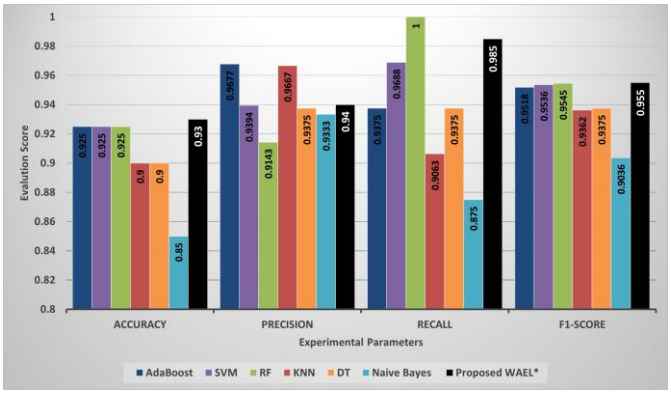

Fig. 6. Predictive performances of the proposed ML models

An EL technique was applied to improve the model performance, leading to a noticeable enhancement in accuracy and recall. Our proposed model achieved the highest accuracy of 93%, as highlighted in Figure 6, demonstrating the effectiveness of the EL method in boosting the model performance.

Table 5 presents the performance scores of the WAEL approach, which uses the top three models: RF, AdaBoost, and SVM. These models were implemented using the Scikit-learn ML library. The EL achieved an accuracy of 93% by effectively increasing the strengths of the individual models. Specifically, the RF model achieved an accuracy of 92.7% with a weight of 0.337, the SVM model had an accuracy of 92.5% with a weight of 0.334, and the AdaBoost model also attained an accuracy of 92.5% with a weight of 0.329. By combining these models using the weighted average method within the EL technique, the resulting accuracy on the test dataset reached 93%, which is considered to be excellent.

Table 5. Weighted average ensemble learning results

Model	Results
RF	Weight: 0.337; Accuracy Score: 92.5%
SVM	Weight: 0.334; Accuracy Score: 92.5%
AdaBoost	Weight: 0.329; Accuracy Score: 92.5%
Proposed WAEL Model*	**93%**

The comparative analysis revealed that RF, SVM, and AdaBoost performed exceptionally well, achieving similar levels of accuracy and F1-scores. In particular, RF obtained the highest recall (100%), indicating its strong ability to correctly identify positive dengue knowledge cases. This superior recall can be attributed to the ensemble nature of RF, which aggregates multiple decision trees to minimize variance and capture complex, non-linear feature interactions. However, such high recall may also indicate a degree of overfitting, especially given the limited dataset size. RF tends to memorize smaller datasets, resulting in nearly perfect classification on training and validation subsets but potentially reduced generalizability to unseen data.

In contrast, SVM and AdaBoost maintained a better balance between precision and recall, reflecting their robustness against noise and overfitting. The proposed WAEL model leverages these complementary strengths of RF's sensitivity to true positives and SVM's boundary generalization to produce a more stable and generalized performance across all evaluation metrics. By combining predictions using optimized weights, WAEL effectively reduces model bias and variance, achieving both high accuracy (93%) and improved recall (98.5%) without overfitting to the training data.

5. CONCLUSION

5.1. Summary of Contributions

This study developed a Weighted Average Ensemble Learning (WAEL) model to enhance dengue outbreak prediction in Bangladesh, with a focus on the Chattogram region. By integrating Support Vector Machine (SVM), Random Forest (RF), and AdaBoost classifiers, the model achieved superior accuracy (93%), precision (94%), and recall (98.5%) compared with standalone models. The research contributes to the field by (i) implementing a rigorous data preprocessing pipeline with MICE-based imputation, (ii) validating the ensemble's stability through cross-validation and confidence intervals, and (iii) identifying key demographic and environmental features such as contact history, residence type, and age as critical determinants of dengue awareness and prevention knowledge. These findings highlight the novelty of combining multiple learners into a context-aware and interpretable framework suitable for resource-constrained public health environments.

5.2. Limitations

Despite its promising performance, this study has several limitations. The dataset comprised only 199 records, which may limit the model's generalizability to larger populations. Although ensemble learning helped mitigate overfitting, a larger and more diverse dataset would provide stronger validation of robustness. Furthermore, the absence of temporal or climatic variables limits the model's ability to capture seasonal outbreak dynamics. Future research should consider incorporating multi-regional data, time-series features, and emerging deep learning architectures such as CNNs, LSTMs, and transformer-based models to improve predictive performance, especially with larger datasets and ensure external validity.

5.3. Policy Relevance

The findings of this study have important implications for dengue management and public health policy in Bangladesh.

The proposed WAEL model, being computationally lightweight and developed with open-source tools, can be readily integrated into existing mobile or web-based surveillance systems used by local health authorities. By connecting with national databases such as those maintained by the Directorate General of Health Services (DGHS), the framework can generate early alerts for high-risk areas based on real-time survey or sensor data. This integration would enable policymakers and health organizations to enhance early warning systems, optimize resource allocation, and implement targeted awareness and prevention initiatives in dengue-prone urban areas.

ACKNOWLEDGEMENT

This study was performed at the American International University-Bangladesh (AIUB). The authors thank the AIUB authority for their support.

REFERENCES

[1] S. Bhatt, P. W. Gething, O. J. Brady, J. P. Messina, A. W. Farlow, C. L. Moyes, et al., "The global distribution and burden of dengue," Nature, vol. 496, no. 7446, pp. 504–507, Apr. 2013, doi: https://doi.org/10.1038/nature12060.

[2] S. Yesmin, S. Sarmin, A. M. Ahammad, M. A. Rafi, and M. J. Hasan, "Epidemiological investigation of the 2019 dengue outbreak in Dhaka, Bangladesh," Journal of Tropical Medicine, vol. 2023, pp. 1–7, 2023, doi: https://doi.org/10.1155/2023/8898453.

[3] M. A. Rob, M. Hossain, M. A. Sattar, I. U. Ahmed, A. F. M. N. Chowdhury, H. M. H. Mehedi, et al., "Circulating dengue virus serotypes, demographics, and epidemiology in the 2023 dengue outbreak in Chittagong, Bangladesh," Eur. J. Microbiol. Immunol., vol. 14, no. 3, pp. 272–279, Aug. 22, 2024, doi: https://doi.org/10.1556/1886.2024.00069.

[4] M. S. Hossain, A. A. Noman, S. M. A. Mamun, A. Al Mosabbir, et al., "Twenty-two years of dengue outbreaks in Bangladesh: epidemiology, clinical spectrum, serotypes, and future disease risks," Trop. Med. Health, vol. 51, art. no. 37, Jul. 11, 2023, doi: https://doi.org/10.1186/s41182-023-00528-6.

[5] J. Clarke, A. Lim, P. Gupte, D. M. Pigott, W. G. Van Panhuis, and O. J. Brady, "A global dataset of publicly available dengue case count data," Scientific Data, vol. 11, no. 1, 2024, doi: https://doi.org/10.1038/s41597-024-03120-7.

[6] M. M. Khan, M. A. H. Miah, M. K. Alam, M. A. Islam, M. A. Rahman, R. I. I. Noor, et al., "Clinico-epidemiological profiling of dengue patients in a non-endemic region of Bangladesh," Trans. R. Soc. Trop. Med. Hyg., vol. 119, no. 1, pp. 58–64, 2024, doi: https://doi.org/10.1093/trstmh/trae074.

[7] M. N. Hasan et al., "The 2023 fatal dengue outbreak in Bangladesh highlights a paradigm shift of geographical distribution of cases," Epidemiology and Infection, vol. 153, 2025, doi: https://doi.org/10.1017/S0950268824001791.

[8] K. M. Y. Arafat, A. Hossain, M. Ikfat, M. A. Amin, K. Tanvir, D. Gomes, and M. Rahman, "FRF-HHO: Early ovarian cancer prediction using explainable fuzzy random forest optimized by Harris Hawks algorithm," Advances in Biomarker Sciences and Technology, vol. 8, 2026, doi: https://doi.org/10.1016/j.abst.2026.01.003.

[9] C. Cortes and V. Vapnik, "Support vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995, doi: https://doi.org/10.1007/BF00994018.

[10] P. Chanprasopchai, I. M. Tang, and P. Pongsumpun, "SIR model for dengue disease with effect of dengue vaccination," Computational and Mathematical Methods in Medicine, vol. 2018, pp. 1–14, 2018, doi: https://doi.org/10.1155/2018/9861572.

[11] M. S. D. P. Nayak and K. A. Narayan, "Forecasting dengue fever incidence using ARIMA analysis," International Journal of Collaborative Research on Internal Medicine & Public Health, vol. 11, no. 6, pp. 924–932, 2019. [Online]. Available: https://www.iomcworld.org/articles/forecasting-dengue-fever-incidence-using-arima-analysis.pdf.

[12] S. G. Kakarla et al., "Lag effect of climatic variables on dengue burden in India," Epidemiology and Infection, vol. 147, 2019, doi: https://doi.org/10.1017/S0950268819000608.

[13] S. Mala and M. Jat, "Geographic information system based spatio temporal dengue fever cluster analysis and mapping," The Egyptian Journal of Remote Sensing and Space Science, vol. 22, no. 8, pp. 297–304, 2019, doi: https://doi.org/10.1016/j.ejrs.2019.08.002.

[14] T. S. Ho et al., "Comparing machine learning with case control models to identify confirmed dengue cases," PLoS Neglected Tropical Diseases, vol. 14, no. 11, p. e0008843, 2020, doi: https://doi.org/10.1371/journal.pntd.0008843.

[15] E. Munarsih and I. Saluza, "Comparison of exponential smoothing method and autoregressive integrated moving average (ARIMA) method in predicting dengue fever cases in the city of Palembang," Journal of Physics: Conference Series, vol. 1521, p. 032100, 2020, doi: https://doi.org/10.1088/1742-6596/1521/3/032100.

[16] W. Hoyos, J. Aguilar, and M. Toro, "Dengue models based on machine learning techniques: A systematic literature review," Artificial Intelligence in Medicine, vol. 119, p. 102157, 2021, doi: https://doi.org/10.1016/j.artmed.2021.102157.

[17] M. Yusuff, "The role of GIS in mapping dengue hotspots," unpublished, 2023.

[18] Sarker and Karim, "Predictive models for dengue outbreak of Dhaka, Bangladesh: Generalized additive models with B spline regression," Jahangirnagar University Journal of Statistical Studies, vol. 37, pp. 101–115, 2023.

[19] M. C. Ramírez Soto, J. V. B. Machuca, D. H. Stalder, D. Champin, and M. G. Martínez Fernández, "SIR SI model with a Gaussian transmission rate: Understanding the dynamics of dengue outbreaks in Lima, Peru," PLoS ONE, vol. 18, no. 4, p. e0284263, 2023, doi: https://doi.org/10.1371/journal.pone.0284263.

[20] K. Tanvir, M. Rahman, and D. Gomes, "A hybrid optimization and Tree-Based learning framework for dengue diagnosis using hematological data," SSRN Electronic Journal, Jan. 2025, doi: https://doi.org/10.2139/ssrn.5687196.

[21] M. J. Hasan et al., "Clinical and epidemiological characteristics of the dengue outbreak of 2024: a multicenter observation from Bangladesh," Tropical Medicine and Health, vol. 53, p. 45, 2025, doi: https://doi.org/10.1186/s41182-025-00691-y.

[22] S. Rahman, F. Mehejabin, and R. Rashid, "Knowledge and prevention practice against dengue vectors among dengue patients and general people in Chattogram, Bangladesh," F1000Research, vol. 11, p. 146, 2022, doi: https://doi.org/10.12688/f1000research.108731.1.

[23] G. Guo, H. Wang, Y. Bell, Davidand Bi, and K. Greer, KNN Model Based Approach in Classification. Berlin, Heidelberg: Springer, 2003.

[24] M. Rahman, M. Hasan, M. M. Billah, and R. J. Sajuti, "Grading system prediction of educational performance analysis using data mining approach," Malaysian Journal of Science and Advanced Technology, vol. 2, no. 4, pp. 204–211, 2022, doi: https://doi.org/10.56532/mjsat.v2i4.96.

[25] A. A. Chowdhury, A. Das, S. K. Saha, M. Rahman, and K. T. Hasan, "Sentiment analysis of COVID 19 vaccination from survey responses in Bangladesh," Research Square, unpublished, 2021.

[26] M. Rahman, M. Hasan, M. M. Billah, and R. J. Sajuti, "Political fake news detection from different news source on social media using machine learning techniques," AIUB Journal of Science and Engineering (AJSE), 2022, doi: https://doi.org/10.53799/ajse.v21i2.383.

[27] P. Favaro and A. Vedaldi, AdaBoost. Boston, MA, USA: Springer US, 2014.

[28] M. A. I. Neloy, N. Nahar, M. S. Hossain, and K. Andersson, "A weighted average ensemble technique to predict heart disease," pp. 17–29, 2022.

[29] N. Dey, M. Srinivas, and R. B. V. Subramanyam, "A novel contactless middle finger knuckle based person identification using ensemble learning," pp. 981–986, 2023.