I. J. Information Technology and Computer Science, **2025, 6, 1-4**
Published Online on December 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2022.06.01

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

Mahfujur Rahman*
American International University-Bangladesh, Dhaka, 1229, Bangladesh
E-mail: mahfuj@aiub.edu
ORCIDiD: https://orcid.org/0000-0002-3245-8767
*CorrespondingAuthor

Noboranjan Dey
American International University-Bangladesh, Dhaka, 1229, Bangladesh
E-mail: noboranjan@aiub.edu
ORCIDiD: https://orcid.org/0009-0007-1991-9198

Nazmus Sakib
American International University-Bangladesh, Dhaka, 1229, Bangladesh
E-mail: 20-43156-1@student.aiub.edu
ORCIDiD: https://orcid.org/0009-0000-8814-5834

Mehedi Hasan
American International University-Bangladesh, Dhaka, 1229, Bangladesh
E-mail: mehedi@aiub.edu
ORCIDiD: https://orcid.org/0000-0002-1146-6274

Abdullah Hel Azmain
American International University-Bangladesh, Dhaka, 1229, Bangladesh
E-mail: 16-32908-3@student.aiub.edu
ORCIDiD: https://orcid.org/0009-0002-2902-2839

Rahul Biswas
American International University-Bangladesh, Dhaka, 1229, Bangladesh
E-mail: rbiswas@aiub.edu
ORCIDiD: https://orcid.org/0009-0009-4459-4756

Received: Date Month, Year; Revised: Date Month, Year; Accepted: Date Month, Year; Published: Date Month, Year

Abstract: Dengue Fever (DF) spreads widely across the globe and poses a serious threat to humanity as one of the major pandemic vector-borne viral diseases. The World Health Organization (WHO) observed that dengue cases have risen sharply worldwide in recent decades. The recent challenge for researchers to address in predicting dengue outbreaks is improving accuracy. Few studies have thoroughly analyzed the risk factors, transmission patterns, and seasonal trends related to dengue outbreak prediction. Several widely used Machine Learning (ML) classification models—namely Adaptive Boosting (AdaBoost), Support Vector Machine (SVM), *k*-Nearest Neighbor (*k*NN), Random Forest (RF), Naïve Bayes (NB), and Decision Tree (DT)—were employed in this study, along with the proposed Ensemble Learning (EL) model. The models were tested and evaluated using a standard dataset collected in Chittagong, Bangladesh. Prior to applying ML models, comprehensive data preprocessing was conducted to ensure efficiency, scalability, trustworthiness, and the reliability of the results in this study. Most dominant ML classification algorithms (AdaBoost, SVM, *k*NN, RF, NB, DT) along with the proposed EL model were applied for performance comparison. Experimental results demonstrated that the proposed Weighted Average Ensemble Learning (WAEL) model increased the accuracy of dengue outbreak predictions to 93%. It can be a valuable tool to fight dengue, enhance awareness, and improve global health outcomes.

IndexTerms: Dengue Fever, Dengue Virus, Dengue Transmission, Machine Learning, Ensemble Learning, Weighted Average Ensemble Learning, Dengue Outbreak Prediction.

1. Introduction

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 infection, with the 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 favorable climatic conditions such as high temperatures, rainfall, and humidity. These factors create ideal breeding environments for mosquitoes, contributing to the disease's cyclical and complex transmission dynamics.

DF has become a growing public health concern in Bangladesh, with urban areas like Chattogram experiencing severe outbreaks. Chattogram, the country's second-largest city, is particularly vulnerable due to rapid urbanization, inadequate waste management, and climate variability. The 2019 outbreak was the worst on record for Bangladesh, with over 100,000 reported cases and 164 confirmed deaths [2]. Chattogram contributed significantly to this national burden, with sharp increases in hospital admissions. In subsequent years, the city continued to experience troubling trends.

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

The clinical presentation of DF is diverse, ranging from mild febrile illness to severe forms such as dengue hemorrhagic 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 and above, males, workers, homemakers, and students, are particularly vulnerable. Additionally, behavioral factors such as smoking and alcohol consumption may contribute to the 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 the disease.

Our study aims to comprehensively analyze dengue's risk factors, transmission patterns, and seasonal trends to address these challenges. In summary, the contributions to this work are mentioned below:

1. Firstly, comprehensive data preprocessing steps have been performed to produce a clean and well-structured dataset optimized for accurate model training and meaningful analytical outcomes.

2. Advanced ML methodologies such as SVM, *k*NN, RF, DT, NB, and AdaBoost are applied to identify high-risk areas, demographic vulnerabilities, and key determinants influencing dengue incidence.

3. Finally, the proposed WAEL technique has been applied, ensuring higher accuracy than any other single model. The performance of the proposed system was compared with state-of-the-art methods. The insights gained from this analysis will contribute to developing targeted public health interventions, strengthening community awareness, and fostering preventive measures to mitigate the health, social, and economic impacts of DF.

The remainder of this paper is structured as follows: Section 2 reviews related literature, highlighting previous studies on dengue risk factors, transmission patterns, and predictive modeling approaches. Section 3 describes the research methodology, including data collection, preprocessing, feature selection, and the ML models used in the study. Section 4 discusses the experimental setup and presents the results of model evaluation. Section 5 provides an in-depth analysis of the results, comparing model performance and identifying key findings. Section 6 summarizes the comparative strengths and weaknesses of the applied methodologies. Finally, Section 7 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 socio-demographic 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 deeply analyzes 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 the previous research studies and their significant findings are presented in Table 1.

Table 1. Summary of studies on dengue forecasting and modeling

Studies	Methods	Key Contributions	Dataset Information	Significance	Limitations
Cortes, C. and Vapnik, V. (1995) [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. (2018) [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, K.A. (2019) [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. (2019) [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. (2019) [13]	Space-time scan statistics, GIS techniques	Spatio-temporal clusters in Delhi, wind-speed association	Daily cases (2010–2012), wind speed data	High-risk cluster detection, resource allocation	The short study period excludes socio-economic factors
Ho, T-S et al. (2020) [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. (2020 [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. (2021) [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. (2023) [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. (2023) [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. (2023) [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. (2024) [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 employed advanced ML algorithms to classify dengue cases and predict outbreaks. Cortes, C. and Vapnik, V. (1995) [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 dependency on kernel functions limit scalability. Ho, T-S et al. (2020) [14] used Deep Neural Networks (DNNs), Logistic Regression (LR), and DTs models on clinical data. In their study, they achieve reliable prediction results using minimal input variables (e.g., age, platelet count). A notable limitation was the exclusion of co-infections (e.g., malaria), which may confound results. Another research performed by Hoyos, W. et al. (2021) [16] applied RF and Geographically Weighted Regression (GWR) methods in their experimentation. They provide robust diagnostic and epidemic modeling, though data quality and language biases in sourcing studies were challenges.

Another primary technique is time series analysis, which has been pivotal in the prediction of outbreaks of dengue. Nayak, M.S.D.P. and Narayan, KA (2019) [11] applied an innovative time series analysis and forecasting method named Seasonal ARIMA. In this study, forecasting monthly dengue cases in Kerala, India, outperforms higher than other ARIMA analyses but relies on secondary data, which may result in risky verification. Another study by Munarsih, E. and Saluza, I. (2020) [15] also applied the ARIMA method and showed superior results with high accuracy and low error matrices. However, manual parameter tuning and ignoring external factors (e.g., climate) reduced generalizability.

Geospatial techniques play an essential role in identifying high-risk zones and transmission patterns. Mala, S. and Jat, M.K. (2019) [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 socio-economic factors, which may have created biases in their outcomes. Yusuff, M. (2023) [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 a significant role in minimizing dengue hazard. Chanprasopchai, P. et al. (2018) [10] and Ramírez-Soto, M.C. et al. (2023) [19] both used SIR and SIR-SI models to evaluate vaccination impacts 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 research. Kakarla, S.G. et al. (2019) [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. (2023) [18] identified humidity and temperature in the Dhaka region as key predictors and applied GAMLSS with B-Spline Regression in their research. Their study uses linear assumptions for precipitation effects, which may oversimplify relationships.

Clarke has performed another research, J. et al. (2024) [20], to standardize global dengue data from 102 countries, enabling high-resolution analyses using the OpenDengue method. Though better public health decision-making outcomes have been found from their research, variability in reporting standards and manual processing requirements posed challenges.

In every ML research, 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. (2020) [14] and Ramírez-Soto, M.C. et al. (2023) [19] which included confirmed cases and climatic variables but often lacked granularity (e.g., serotype details). Kakarla, S.G. et al. (2019) [12] and Sarker, I. & Karim, M.R. (2023) [18] integrate epidemiological and environmental climate records but exclude urban-specific factors (e.g., sanitation). Although ethical and technical barriers hindered scalability, Yusuff, M. (2023) [17] combines multi-source geospatial health records with satellite imagery.

Recent advancements in ML and computational modeling have significantly enhanced dengue prediction and control efforts. ML techniques, such as SVNs and DNNs, have improved outbreak forecasting accuracy and optimized resource allocation. Spatial tools like GIS-based hotspot analysis have enabled targeted interventions by identifying high-risk zones. At the same time, global databases such as OpenDengue have facilitated cross-country comparisons and standardized data analysis. However, several challenges persist, including data limitations such as short study durations, underreporting, and reliance on secondary data. Model constraints, such as oversimplified assumptions in compartmental models (e.g., SIR frameworks) and the exclusion of confounding variables (e.g., co-infections), reduce generalizability. Additionally, ethical concerns in geospatial studies and the high technical expertise required for ML implementation pose barriers to scalability. Addressing these limitations through multi-modal data integration, federated learning, and standardized reporting protocols will be crucial for advancing dengue research and public health strategies.

This review underscores the transformative potential of ML in advancing dengue management strategies. Addressing these critical gaps will refine predictive models and strengthen global dengue surveillance and intervention efforts, ultimately contributing to more effective public health outcomes. By adopting these forward-looking approaches, researchers can harness the full power of ML to mitigate the burden of dengue worldwide.

3. Research Methodology

DENV, a mosquito-borne disease, represents a significant global public health concern due to its extensive economic, social, and health-related impacts. This study employs 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, the 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 was retrieved from the Figshare data repository titled “Knowledge and prevention practice regarding dengue fever” and contains comprehensive information regarding demographic characteristics, knowledge and awareness, prevention practices, environmental factors, and health outcomes. The summary of the dataset is mentioned in Table 2.

Table 2. Dengue virus awareness and prevention techniques dataset summary

Dataset Name	**Data Source**	**Number of Attributes**	**Number 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. The survey was conducted among residents of Chattogram, Bangladesh. The dataset includes 300 entries and 41 variables, encompassing both categorical and numerical data. It is structured to facilitate the analysis of factors influencing knowledge and preventive practices related to dengue. Key demographic attributes captured in the dataset include 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.

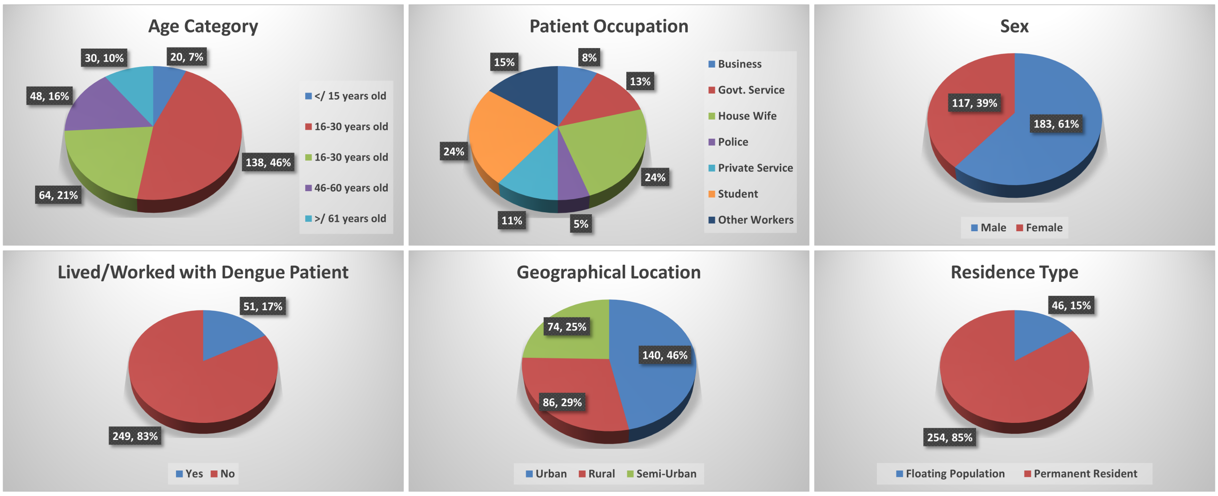

Fig 1. The visual distribution of different categorical attributes of the dataset

3.2. Data Preprocessing

Data preprocessing is a critical component of the research process, aimed at ensuring the quality and suitability of data for subsequent analysis. This phase encompasses several tasks, including data integration, selection, cleaning, transformation, feature selection, and feature engineering. These procedures are essential for converting raw data into a structured and analyzable format. Effective preprocessing enhances the interpretability and usability of the dataset, facilitates the extraction of relevant features, and supports the accurate application of predictive modeling techniques. Figure 2 illustrates the data preprocessing workflow and its implementation within the context of this study.

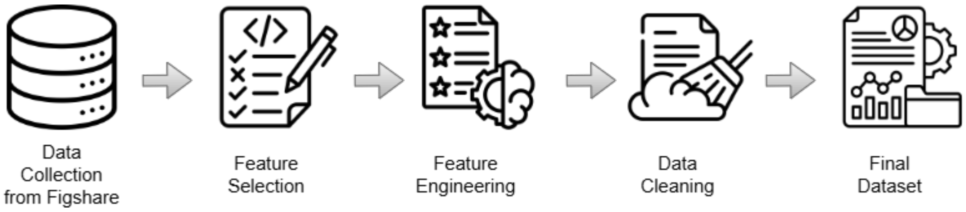

Fig 2. Data preprocessing techniques

3.2.1 Feature Selection

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

This study's collected dataset comprises 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 9 attributes across 300 instances, forming the basis for the subsequent analysis and model development phases.

3.2.2 Feature Engineering

Feature engineering is the process of transforming raw data into a more informative and structured format to improve the performance and accuracy of ML models. The primary objective of this step is to ensure data integrity, preserve statistical power, and prevent biased or inaccurate predictions, particularly by addressing issues such as missing data. Missing values is a common challenge in real-world datasets, arising when specific variables lack corresponding data points. Such incomplete information can adversely affect analytical models' accuracy, reliability, and generalizability. Therefore, effective handling of missing data is essential to ensure robust and unbiased results.

This study examined the dataset comprising primary data collected from a survey related to dengue case studies with 300 instances of missing values. Several instances with incomplete data were identified. To maintain data quality and integrity, two instances containing multiple missing attributes were manually removed. Subsequently, the Multiple Imputation by Chained Equations (MICE) method was employed to address 47 missing values through probabilistic imputation. Following this process, the refined dataset consisted of 298 instances, each with 9 attributes, and was used for further analysis and model development.

3.2.3 Data Cleaning

Data cleaning represents a foundational and critical step in ML research, a prerequisite for reliable data analysis. It involves the identification and 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 redundancy and correcting inconsistencies.

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

After successfully completing data preprocessing, the final dataset comprises 199 instances and 9 attributes. This refined dataset is free from missing values, redundant attributes, and duplicate entries, thereby ensuring its quality and suitability for analysis. The binary classification for the target variable is the Total Knowledge Score derived from the survey responses. These comprehensive preprocessing steps have 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 further analysis.

3.2.4 Dataset Splitting

Dataset splitting is a crucial step in ML that involves partitioning the data into training, validation, and testing subsets. This process is essential for training the model, tuning hyperparameters, and evaluating performance, thereby helping to prevent overfitting and ensuring the model's ability to generalize to unseen data. The primary objective of dataset splitting is to promote model generalization and avoid overfitting.

This study's dataset comprises 199 instances with 9 attributes following data preprocessing. To optimize model performance, the data was split using a structured approach. Initially, 80% of the dataset, equivalent to 159 instances, was allocated for model training. Within this subset, 75% (representing 60% of the total dataset, approximately 119 instances) was used for training the model, while the remaining 25% (20% of the total dataset, approximately 40 instances) was reserved for validation. The final 20% of the overall dataset, consisting of 40 instances, was held out for testing to assess the model's general performance.

This allocation ensures a balanced and systematic data distribution [50], with 60% used for training, 20% for validation, and 20% for testing, providing a reliable framework for robust model development and evaluation. The dataset-splitting strategy employed in this study is illustrated in Figure 3.

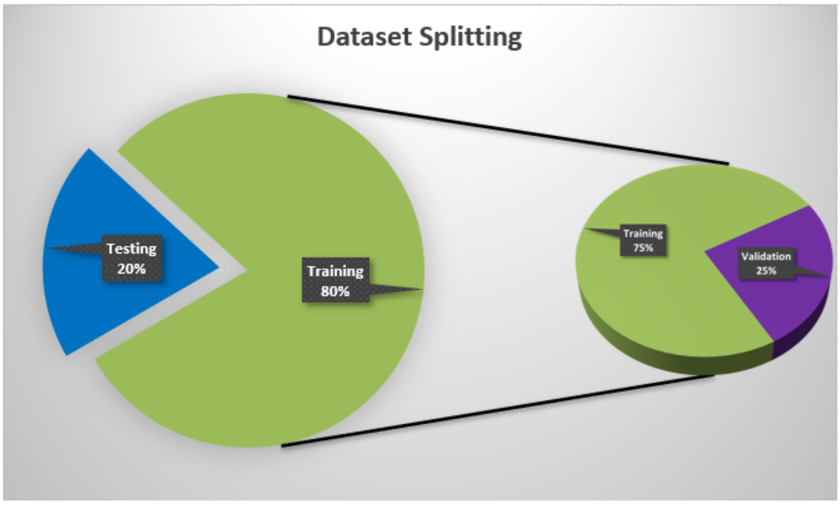

Fig 3. The visualization of dataset splitting

3.3. Proposed Methodology

The research aims 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 collecting raw data, where data is gathered for analysis. This step involves refining the dataset by selecting important features, removing unnecessary attributes, addressing missing values, and eliminating duplicate data to enhance model performance. Once the data was preprocessed, widely used ML models were chosen, such as DT, SVM, *k*NN, AdaBoost, and RF, and their performances were tested for dengue pattern transmission. Following evaluation, the three best-performing classification algorithms are selected based on the F-measure. The selected classifiers design a novel EL classification approach using a weighted average mechanism. The following subsections provide a detailed explanation of the classification algorithms and weighted average mechanisms.

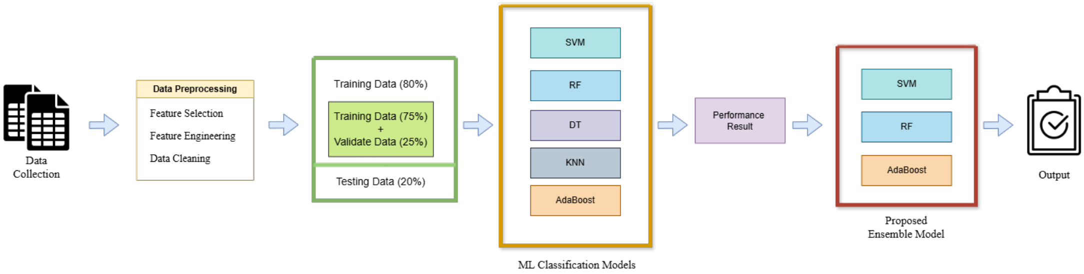

Fig 4. Overall system architecture for dengue prediction

3.3.1 Classification Methods

A. **Support Vector Machine (SVM)**
The SVM is a supervised learning algorithm for classification tasks. It identifies the optimal hyperplane that separates data points into distinct classes while maximizing the margin between them [22]. This research implements SVM to classify the Total Knowledge Score as a target attribute. The model achieved commendable accuracy and precision, demonstrating its potential for predictive analysis.

B. ***k*-Nearest Neighbor (*k*NN)**
The *k*NN 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, *k*NN delivered high precision and recall in our analysis. The model performed well on the dataset, particularly with optimized '*k*' values.

C. **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 tasks. In this research, 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 made it one of the most reliable standalone models.

D. **Naïve Bayes (NB)**
NB is a probabilistic classifier based on Bayes' Theorem, assuming feature independence [25]. In our research, it served as a baseline model, showing moderate accuracy and providing valuable comparative insights. Despite its simplicity, the algorithm's predictive capabilities were constrained by the interdependencies between features in the dataset.

E. **Decision Tree (DT)**
DTs are intuitive models that split data based on feature values to make predictions [26]. This algorithm delivered solid accuracy and interpretability, providing foundational insights into key decision paths in the dataset. However, DTs tended to overfit, particularly with noisy or imbalanced data.

F. **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 accuracy and precision. Its iterative weighting of challenging data points proved effective, although sensitivity to outliers occasionally impacted its results.

G. **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 a ML approach combining 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 a single predictive model. The proposed WAEL model assigns specific weights to each model's output, with the final prediction derived from their weighted average.

$$\hat{y} = \sum_{i=1}^{n} w_i \cdot \widehat{y_i} \tag{1}$$

Where, $\hat{y}$ is the final EL prediction, n is the number of models, w_i is the weight assigned to the i^{th} model, and $\widehat{y_i}$ is the prediction from the i^{th} individual model. In our proposed model, we have used three ML models. Figure 5 visually represents our proposed WAEL model with technical details.

$$\hat{y} = w_1 \cdot \widehat{y_1} + w_2 \cdot \widehat{y_2} + w_3 \cdot \widehat{y_3} \tag{2}$$

With the constraint: $w_1 + w_2 + w_3 = 1$

4. Experimental Results

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

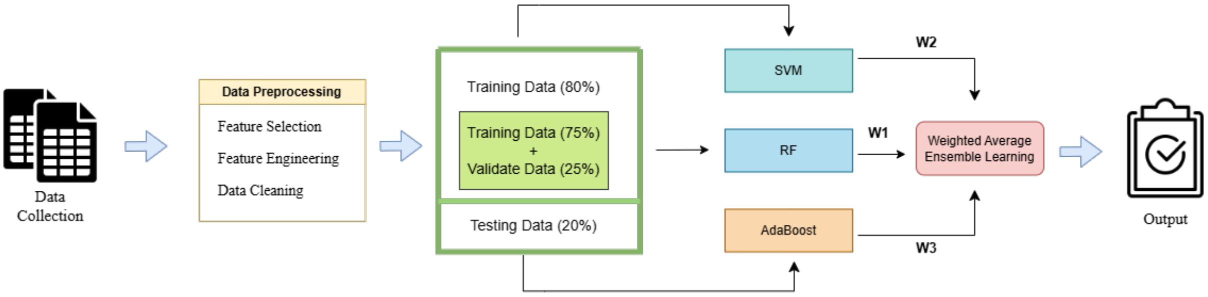

Fig 5. Proposed weighted average ensemble learning model

4.1. Result Analysis

The performance analysis of various ML algorithms revealed key insights into their suitability for predicting an individual's knowledge about dengue disease from input data. This section elaborates on the results, comparing the strengths and limitations of each model to provide a comprehensive understanding of their effectiveness. Table 3 illustrates the performance metrics of various ML models used in this study, including accuracy, precision, recall, and F1 score.

Table 3. 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 their F1 scores ranging from 0.9518 to 0.9545. *k*NN 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 performance remained relatively similar.

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

Table 4. Weighted average ensemble learning results

Model	Result
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%**

Table 4 presents the performance scores of the WAEL approach, which uses the top three models: RF, SVM, and AdaBoost. 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 an excellent result for this problem.

5. Conclusion

Dengue is a common viral fever and is also considered a life-threatening disease. As mentioned in the introduction, our primary purpose was to identify high-risk regions, vulnerable demographic groups, and significant factors affecting dengue incidence. RF emerged as the most reliable model among standalone algorithms, demonstrating strong recall and

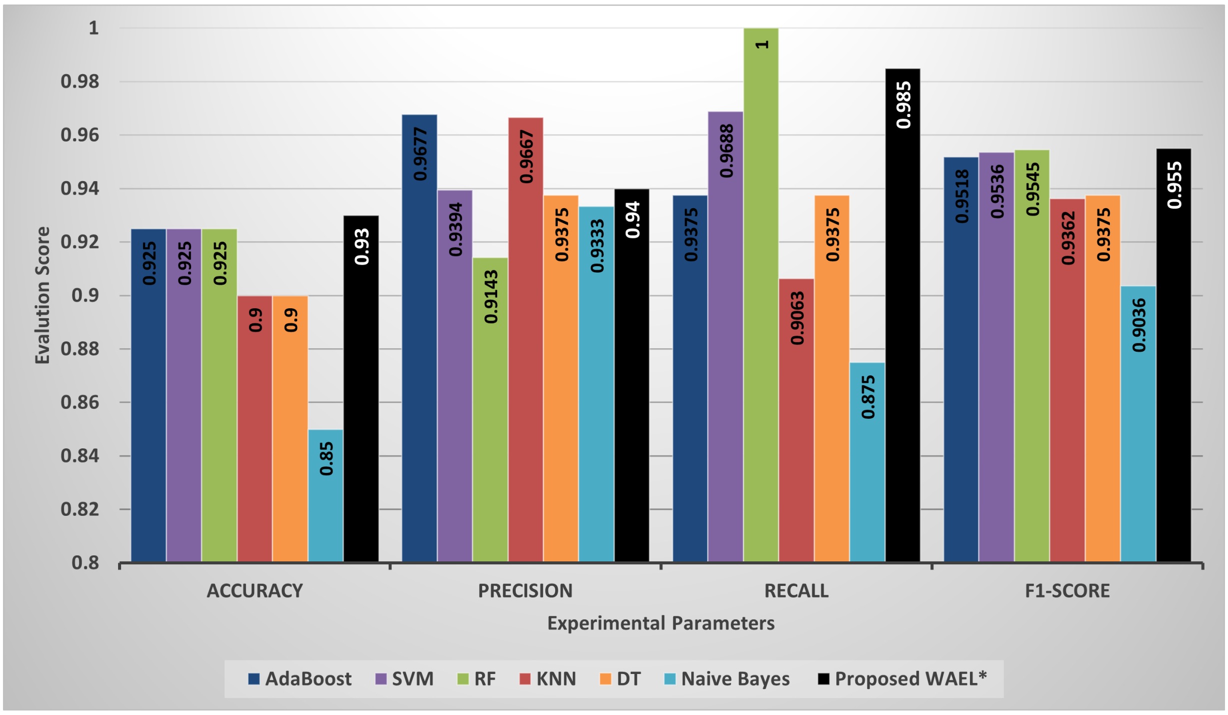

Fig 6. Predictive performances of the proposed ML models

an impressive F1 score. Its ability to handle complex patterns and diverse features contributed to its superior performance. AdaBoost also showcased robust results, achieving high accuracy and an excellent F1 score, further highlighting its effectiveness in dengue prediction. However, our proposed WAEL model outperformed all individual models by achieving high recall and improved accuracy. These scores indicate that our model is the most reliable approach, effectively leveraging the complementary strengths of its base models. Additionally, a comparative analysis with state-of-the-art models validated the superiority of our proposed approach. The insights generated from our study can serve as a foundation for empowering individuals and communities to take proactive measures against dengue transmission. Our research provides a strong foundation for further studies predicting dengue transmission patterns. In addition to improving predictive capabilities, our research aims to increase public awareness about dengue prevention and control. Health authorities and policymakers can design targeted educational campaigns and community outreach programs by identifying the key factors influencing individuals' knowledge about dengue disease. Future work can explore larger datasets, incorporate additional attributes, and apply advanced deep learning methods to refine predictions and broaden the scope of insights. Through these efforts, predictive modeling can be a valuable tool in combating dengue, raising awareness, and improving global health outcomes.

Data Availability

The dataset that support the findings of this study are publicly accessible on Figshare and available at the following link: *Knowledge and prevention practice regarding dengue fever.*

Competing Interest

The authors declare that no known financial conflicts of interest or personal relationships could have influenced the outcome of this work.

Acknowledgment

This work has been performed/carried out in AIUB. The authors thank the AIUB authority for their financial support and any other support.

Funding

This work didn't receive any external funding.

References

[1] World Health Organization: WHO and World Health Organization: WHO. dengue and severe dengue. April 2024.

[2] Sabrina Yesmin, Shahnoor Sarmin, Alamgir Mustak Ahammad, Md. Abdur Rafi, and Mohammad Jahid Hasan. Epidemiological investigation of the 2019 dengue outbreak in dhaka, bangladesh. *Journal of Tropical Medicine*, 2023:1–7, March 2023.

[3] Tbs Report. Dengue death toll breaks record in chattogram. *The Business Standard*, august 2023.

[4] Nasir Uddin Rocky. Dengue cases on the rise in chittagong. *Dhaka Tribune*, june 2023.

[5] Tbs Report. Number of dengue patients in ctg increased ten fold compared to last year. *The Business Standard*, may 2023.

[6] Star Digital Report. Dengue: 4 die, 239 hospitalised in a day. *The Daily Star*, December 2023.

[7] Mohammad Nayeem Hasan, Mahbubur Rahman, Meraj Uddin, Shah Ali Akbar Ashrafi, Kazi Mizanur Rahman, Kishor Kumar Paul, Mohammad Ferdous Rahman Sarker, Farhana Haque, Avinash Sharma, Danai Papakonstantinou, Priyamvada Paudyal, Md Asaduzzaman, Alimuddin Zumla, and Najmul Haider. The 2023 fatal dengue outbreak in bangladesh highlights a paradigm shift of geographical distribution of cases. *Epidemiology and Infection*, 153, January 2025.

[8] Tbs Report. Dengue cases cross 10,000 mark since jan. *The Business Standard*, July 2023.

[9] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995.

[10] Pratchaya Chanprasopchai, I. Ming Tang, and Puntani Pongsumpun. Sir model for dengue disease with effect of dengue vaccination. *Computational and Mathematical Methods in Medicine*, 2018:1–14, August 2018.

[11] M Siva Durga Prasad Nayak and KA Narayan. Forecasting dengue fever incidence using arima analysis. *International Journal of Collaborative Research on Internal Medicine & Public Health*, 11(6):924–932, January 2019.

[12] Satya Ganesh Kakarla, Cyril Caminade, Srinivasa Rao Mutheneni, Andrew P Morse, Suryanaryana Murty Upadhyayula, Madhusudhan Rao Kadiri, and Sriram Kumaraswamy. Lag effect of climatic variables on dengue burden in india. *Epidemiology and Infection*, 147, January 2019.

[13] Shuchi Mala and Mahesh Jat. Geographic information system based spatio-temporal dengue fever cluster analysis and mapping. *The Egyptian Journal of Remote Sensing and Space Science*, 22:297–304, 08 2019.

[14] Tzong-Shiann Ho, Ting-Chia Weng, Jung-Der Wang, Hsieh-Cheng Han, Hao-Chien Cheng, Chun-Chieh Yang, Chih-Hen Yu, Yen-Jung Liu, Chien Hsiang Hu, Chun-Yu Huang, Ming-Hong Chen, Chwan-Chuen King, Yen-Jen Oyang, and Ching-Chuan Liu. Comparing machine learning with case-control models to identify confirmed dengue cases. *PLoS Neglected Tropical Diseases*, 14(11):e0008843, November 2020.

[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*, 1521(3):032100, April 2020.

[16] William Hoyos, Jose Aguilar, and Mauricio Toro. Dengue models based on machine learning techniques: A systematic literature review. *Artificial Intelligence in Medicine*, 119:102157, August 2021.

[17] Mariam Yusuff. The role of gis in mapping dengue hotspots. 06 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*, 37:101–115, 2023.

[19] Max Carlos Ramírez-Soto, Juan Vicente Bogado Machuca, Diego H. Stalder, Denisse Champin, Maria G. Mártinez-Fernández, and Christian E. Schaerer. Sir-si model with a gaussian transmission rate: Understanding the dynamics of dengue outbreaks in lima, peru. *PLoS ONE*, 18(4):e0284263, April 2023.

[20] 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*, 11(1), March 2024.

[21] Sahidur Rahman, Fatema Mehejabin, and Rumana Rashid. Knowledge and prevention practice against dengue vectors among dengue patients and general people in chattogram, bangladesh. *F1000Research*, 11:146, February 2022.

[22] C. Cortes and V Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, September 1995.

[23] Gongde Guo, Hui Wang, Yaxin Bell, Davidand Bi, and Kieran Greer. *KNN Model-Based Approach in Classification*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[24] Mahfujur Rahman, Mehedi Hasan, Md Masum Billah, and Rukaiya Jahan Sajuti. Grading system prediction of educational performance analysis using data mining approach. *Malaysian Journal of Science and Advanced Technology*, 2(4):204–211, Nov. 2022.

[25] Anjir Ahmed Chowdhury, Argho Das, Suben Kumer Saha, Mahfujur Rahman, and Khandaker Tabin Hasan. Sentiment analysis of covid-19 vaccination from survey responses in bangladesh. *Research Square*, 2021.

[26] Mahfujur Rahman, Mehedi Hasan, Md Masum Billah, and Rukaiya Jahan Sajuti. Political fake news detection from different news source on social media using machine learning techniques. *AIUB Journal of Science and Engineering (AJSE)*, 2022.

[27] Paolo Favaro and Andrea Vedaldi. *AdaBoost*. Springer US, Boston, MA, 2014.

[28] Md. Arif Istiek Neloy, Nazmun Nahar, Mohammad Shahadat Hossain, and Karl Andersson. A weighted average ensemble technique to predict heart disease. pages 17–29, 2022.

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

Authors' Profiles

Mahfujur Rahman received his B.Sc. in Computer Science and Engineering and M.Sc. in Intelligent Systems from American International University-Bangladesh (AIUB). He began his academic career at AIUB, where he has been serving since 2020, initially as a Lecturer and currently as an Assistant Professor in the Department of Computer Science. His primary research interests include Intelligent Systems, Natural Language Processing, Machine Learning, Data Science, Medical Image Processing, the Internet of Things, and Big Data Analytics. He has over five years of experience supervising undergraduate thesis projects and has contributed as a reviewer for several national and international peer-reviewed journals and conferences. He can be contacted by email: mahfuj@aiub.edu

Noboranjan Dey is a lecturer in the Department of Computer Science and Engineering at American International University-Bangladesh (AIUB). He hols a Master's degree (M.Tech) in Computer Science & Engineering from National Institute of Technology Warangal, India and a Bachelor of Science degree (B.Sc.) in Computer Science & Engineering from American International University -Bangladesh (AIUB). His research interests include Deep Learning, Neural Networks, Computer Vision, Image Processing, Medical Imaging, Biometrics. He can be contacted by email: noboranjan@aiub.edu

Nazmus Sakib is a Computer Science graduate from American International University-Bangladesh. His primary research interests lie in the fields of machine learning algorithms and web development, where he is passionate about exploring innovative solutions and building intelligent, scalable systems. He can be contacted by email: 20-43156-1@student.aiub.edu

Mehedi Hasan is currently working as a senior assistant professor in the Faculty of Engineering of American International University-Bangladesh (AIUB). He currently pursuing his PhD from a reputed university, BUET in field of Nanotechnology and solid-state devices. He has received his MS degree from EEE Department of American International University-Bangladesh(AIUB) in Bangladesh, Bangladesh in 2015. He received his Bachelor Degree from EEE Department of American International University-Bangladesh(AIUB) in Bangladesh, Bangladesh in 2013. He worked as a Teaching assistant in AIUB in EEE & COE Department from Sep, 2013 to Jan, 2015. He achieved SUMMA CUM LAUDE and EWU Math Olympiad award. His research interests Nanotechnology, Qauntum wire, Quantum Dot, Nanoparticle, Nanotransistor, Spintronics, nano-photonics, Nanosensor, Band Structure Model, Compound semiconductor, Meta Materials, VLSI & ULSI Design. He can be contacted by email: mehedi@aiub.edu

Abdullah Hel Azmain received his B.Sc. in Computer Science and Engineering from the American International University - Bangladesh(AIUB), Dhaka, Bangladesh in 2021. He is currently working as a software engineer at Sulacotec, Dhaka, Bangladesh. Abdullah has an enthusiasm over the reaches field of Artificial Intelligence, Data Mining and Software Engineering. He can be contacted by email: 16-32908-3@student.aiub.edu

Rahul Biswas currently serving as a Lecturer in the Computer Science Department under the Faculty of Science and Technology at American International University-Bangladesh (AIUB). In 2021, he completed his Bachelor's in Computer Science and Engineering, focusing on Information Security, from the Vellore Institute of Technology (VIT) in Vellore, India. Following my insatiable curiosity, he pursued a Master's degree at the respected Indian Institute of Technology Ropar (IIT Ropar) and completed it in 2023. During my academic journey, he worked on a significant government project at IIT Ropar called iHub-AWaDH (Agriculture and Water Technology Development Hub) and contributed to agriculture technology development. He also had the opportunity to intern at CTG IT Service Limited, where he worked as a web developer and software tester. Beyond his local experiences, he have been involved in international conferences as a volunteer and have positively impacted the global academic stage. His research interests encompass Natural Language Processing (NLP), Applied Machine Learning, Deep Learning, CNN, Agriculture Data Analysis and Collection. He can be contacted by email: rbiswas@aiub.edu

”Enhancing Dengue Outbreak Prediction in Bangladesh: A Weighted Average Ensemble Machine Learning Approach”, International Journal of Information Technology and Computer Science(IJITCS), Vol.14, No.6, pp.1-4, 2022. DOI:10.5815/ijitcs.2022.06.01