Dengue Prediction Using Ensemble Machine Learning Models

A. K. M. Masudur Rahman[1], MST. Sumya Jafrin[1]
[1]Dept. of Computer Science and Engineering, Bangladesh Army University of Science and Technology (BAUST), Saidpur, Bangladesh
Emails: akmmasudurrahmangaurab@gmail.com, sumya6577jafrin@gmail.com

***Abstract*—Dengue fever is a mosquito-borne viral infection that poses serious public health risks in tropical regions, particularly in Bangladesh where recent outbreaks have caused significant morbidity and mortality. while traditional diagnoses are reliable, they often take too long during peak outbreaks due to limited resources and testing capacity. This research demonstrates a machine-learning method for early dengue prediction that uses demographic, clinical, and environmental data. We trained and evaluated Nine supervised machine learning models such as Logistic Regression, Gaussian Naïve Bayes, Decision Tree, K-Nearest Neighbors, Support Vector Machine, Random Forest, XGBoost, Gradient Boosting, CatBoost, and LightGBM and created a soft-voting ensemble model to combine all these base classifiers, which improved the reliability and predictive accuracy. The results indicate that Logistic Regression has achieved the highest individual accuracy of 0.96, while the ensemble model maintained a cross-validation accuracy of 0.95. The proposed framework displays the potential of ensemble learning as a quick, reliable, and scalable tool to support clinical decision-making and public health efforts during dengue outbreaks in areas with limited resources.**

***Keywords*—Dengue prediction, ensemble learning, machine learning, clinical decision support, supervised algorithms.**

I. Introduction

Dengue, an endemic arboviral disease caused by infection with one of four closely related viral serotypes such as DENV-1, DENV-2, DENV-3, and DENV-4 [1] has emerged as the new global health threat, especially in South-East Asia. In Bangladesh, it has escalated to a substantial public health concern. The country reported 321,179 confirmed cases and 1,705 deaths in 2023, marking it as the deadliest dengue outbreak in the history of Bangladesh [1].

Dengue has become a significant threat in all 64 districts of Bangladesh, but recent outbreaks have exhibited a new phenomena, with cases increasing both inside and outside Dhaka. This marks Dhaka as the hotspot for this viral infection [1]. The case-fatality ratio (CFR) during the 2023 outbreak had reached approximately 0.53%, which indicates that this affects older adults and women disproportionately [2]. On top of that, studies conducted in 2023 revealed that early and irregular seasonality which is linked to unusually high rainfall, temperature, and humidity, accelerates Aedes mosquito proliferation [3].

In Bangladesh, dengue transmission is carried out primarily by the Aedes aegypti mosquito, with Aedes albopictus acting as a secondary vector. Environmental and demographic conditions, such as monsoon-driven rainfall, rapid urbanization, and high population density, further accelerates vector breeding and virus transmission [1], [2]. Dengue infection can range from mild feverish illness to life-threatening severe dengue, including hemorrhage, organ impairment, and plasma leakage. Records of the 2023 outbreak shows that a considerable number of fatal cases developed dengue shock syndrome (DSS) [2].

Although traditional laboratory diagnostics such as NS1 antigen detection and IgG/IgM serological inspections still remain essential for dengue confirmation, they often face delays during major outbreaks due to limited capacity, lack of infrastructures, and logistical constraints. Therefore, the need for early-warning systems and predictive tools became crucial for timely clinical and public health interventions.

Machine learning algorithms and models provide a data-driven approach by analyzing demographic, clinical, and environmental attributes to dengue prediction. Ensemble learning methods, which combine multiple classifiers to achieve better real-time performance, can improve the robustness of the model and the predictive accuracy in complex public health and disease-related datasets from the real-world epidemics [4], [5]. Several previous studies have showed the application of machine learning and ensemble techniques for dengue prediction, including ensemble neural networks using climatic variables [4].

In this study, we propose an ensemble machine-learning framework for dengue prediction in Bangladesh which we believe will play a significant role in reducing the delays in dengue confirmation during major outbreaks and will help in dealing this phenomenon. Nine supervised machine learning models such as Logistic Regression, Gaussian Naïve Bayes, Decision Tree, K-Nearest Neighbors, Support Vector Machine, Random Forest, XGBoost, Gradient Boosting, CatBoost, and LightGBM, were trained and their real-time performences were compared. A soft-voting ensemble model was implemented to integrate all base classifiers. By using both geographically and clinically informed data, the proposed system aims to provide fast, reliable, and scalable predictions, supporting outbreak management and clinical decision-making in resource-constrained environments. The major contributions of this study are given below:

1) **Dataset Development:** Patient reports were collected from Dhaka and data augmentation techniques were

applied to create a large, diverse dataset and prevent model overfitting.

2) **Model Design:** Nine supervised machine learning models such as Logistic Regression, Gaussian Naïve Bayes, Decision Tree, K-Nearest Neighbors, Support Vector Machine, Random Forest, XGBoost, Gradient Boosting, CatBoost, and LightGBM, were trained and optimized for dengue prediction.
3) **Ensemble Methodology:** A stacking ensemble model was developed using the above nine base classifiers with XGBoost as the meta-learner. Threshold tuning was applied to maximize prediction accuracy.
4) **Evaluation and Performance:** The models were evaluated using accuracy, ROC AUC, precision, recall, F1 score, and confusion matrices. The stacking ensemble achieved high accuracy and robust predictions, outperforming individual models.

II. Literature Review

Dengue fever has emerged as a major public health concern in tropical and subtropical regions, including Bangladesh, necessitating timely prediction systems for outbreak mitigation. In recent years, numerous studies have explored machine learning (ML) techniques for dengue prediction, leveraging demographic, clinical, and environmental features. This section critically reviews existing literature and positions the present study within this research landscape.

Islam et al. [7] developed a predictive framework using hospitalized patient data combined with meteorological and socioeconomic features across 11 districts in Bangladesh. They employed Multiple Linear Regression (MLR) and Support Vector Regression (SVR), achieving moderate accuracies of 67% and 75%, respectively. While the study demonstrated the feasibility of ML for dengue prediction, the reliance on only two models limited the ability to capture complex non-linear interactions in clinical and demographic data. Furthermore, the framework primarily addressed aggregated predictions rather than patient-level outcomes, reducing its applicability for early detection in clinical settings.

Mobin [6] proposed a multivariate forecasting approach using Bayesian downscaling to convert monthly dengue counts into daily estimates. Integrating high-resolution temporal data improved forecasting performance; however, the study focused predominantly on population-level temporal trends rather than individualized risk prediction. Consequently, while effective for public health surveillance, the model was not optimized for real-time clinical decision support at the patient level.

Hossain et al. [9] implemented Random Forest, XGBoost, and LightGBM models using sociodemographic, climatic, and landscape features. Their ensemble approach, supplemented by SHAP-based interpretability, captured non-linear effects and provided higher predictive performance. Nevertheless, the study concentrated on aggregated population risk, lacking patient-specific predictive capabilities. Additionally, the reliance on environmental variables limited the model's ability to leverage serological or clinical markers for early diagnosis.

Liu et al. [8] conducted a comparative evaluation of SARIMA, Multi-Layer Perceptron (MLP), XGBoost, and SVR models using monthly dengue surveillance data. Their analysis highlighted the superiority of ML approaches over classical statistical models in capturing non-linear relationships. However, the limited temporal resolution and absence of ensemble methods restricted robustness, and patient-level prediction was not addressed.

Braga et al. [10] implemented an ensemble neural network framework in Brazil, combining spatial and temporal predictors for dengue incidence forecasting. Although the ensemble achieved high long-term predictive accuracy, the model was computationally intensive and designed for population-level prediction, making it less suitable for real-time clinical deployment. Similarly, Panja et al. [4] introduced a wavelet-neural network ensemble (XEWNet) to model non-linear climatic effects. The approach demonstrated superior long-term forecasting but required large-scale datasets and substantial computational resources, limiting transferability to patient-level clinical settings.

Ferdousi et al. [11] introduced a windowed correlation-based feature selection strategy for RNN-based dengue prediction. By incorporating spatially adjacent incidence data and time-shifted correlations, the model achieved improved accuracy for temporal forecasts. Nevertheless, reliance on spatially dense datasets restricted generalizability to regions with sparse reporting, and patient-level prediction remained unaddressed.

A hybrid epidemiological-ML approach [12] combined sinusoidal modeling to capture seasonal transmission dynamics with Prophet forecasting. While effective in identifying outbreak peaks and providing insights into contact rate variations, the model focused on aggregated trends rather than individual patient risk, limiting its utility for clinical decision support.

In contrast to these studies, the present work integrates serological markers (NS1, IgG, IgM), demographic attributes, and environmental features to perform patient-level dengue prediction. We employ a soft-voting ensemble of seven supervised classifiers, including linear, probabilistic, tree-based, and boosting models, to leverage complementary strengths. This design captures both linear and non-linear relationships, enhances predictive accuracy, and minimizes false negatives, which is critical for early detection. Moreover, by focusing on individual patient outcomes, our framework addresses the limitations of prior population-level models, offers robustness against overfitting, and provides a scalable, interpretable solution suitable for real-time clinical decision support in dengue-endemic regions such as Bangladesh.

III. Materials and Methods

A. Dataset and Preprocessing

The dataset used in this research consists of 1,000 patient cases that were collected from Dhaka, Bangladesh, which contains demographic, serological, and environmental attributes, like *Gender, Age, NS1, IgG, IgM, Area, AreaType, HouseType, District,* and *Outcome*. The *Outcome* contains two classes: 0

(Dengue Negative) and 1 (Dengue Positive). Each row in this dataset represents an individual patient case.

In order to prevent overfitting and to enhance model performance, data augmentation was performed, which effectively doubled the dataset to 2,000 patient cases. Exploratory Data Analysis (EDA) was performed to understand feature distributions, relationships, and potential class imbalances in the dataset, and feature engineering was performed to create meaningful new features that could enhance predictive performance of the models.

Categorical variables such as *Gender, Area, AreaType, HouseType,* and *District* were transformed using both One-Hot Encoding and Label Encoding, which ensured better model fitting and performance. Label Encoding was particularly chosen for tree-based models like Random Forest, XGBoost, CatBoost, and LightGBM, which can handle integer-encoded categorical variables without introducing bias. Numerical features, like *Age, NS1, IgG,* and *IgM*, were standardized using `StandardScaler` to zero mean and unit variance, which ensured distance- and margin-based models, such as KNN and SVM, performed effectively.

Correlation analysis was used to study feature interactions, and a correlation heatmap was generated. Class distribution of the *Outcome* classes was visualized using a countplot to determine the balance between dengue-positive and dengue-negative cases.

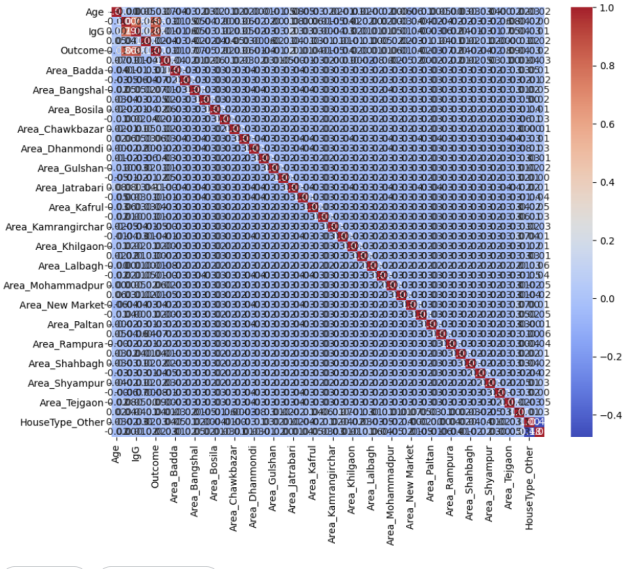

Fig. 1. Heatmap of Feature Correlations

```
<Axes: xlabel='Outcome', ylabel='count'>
```

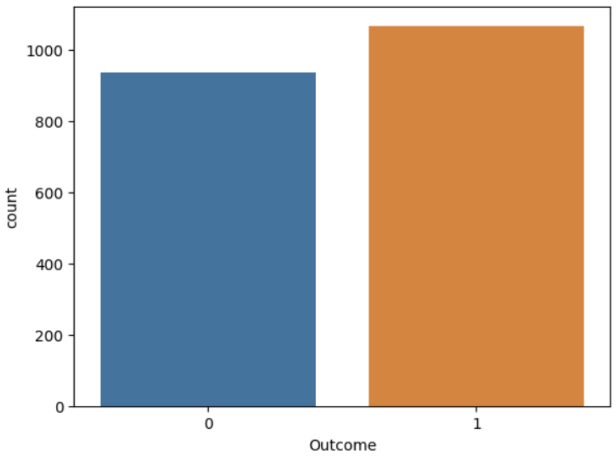

Fig. 2. Class Distribution of the Outcome classes using Countplot

1) Train-Test Split and Data Processing: After preprocessing, the dataset containing 2,000 samples was split into training and testing sets using an 80:20 ratio, with 1,600 samples for training and 400 samples for testing. Stratified splitting ensured that the proportion of dengue-positive and dengue-negative cases was maintained in both sets, preserving class balance and ensuring proper model evaluation.

Feature scaling and feature encoding as described above were applied to the training data, while the test set was transformed using the same scalers and encoders fitted on the training data. This prevented data leakage and provided an unbiased evaluation of predictive performance.

B. Machine Learning Algorithms

1) Logistic Regression: Logistic Regression (LR) is a probabilistic model for binary classification. It models the probability that a given input vector $\mathbf{x} = [x_1, x_2, ..., x_n]$ belongs to the positive class as:

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x} + b)}} \tag{1}$$

where $\mathbf{w}$ is the vector of feature coefficients, and b is the bias term. The model is trained by maximizing the log-likelihood:

$$J(\theta) = -\ell(\theta) = -\sum_{i=1}^{n} \Big[y_i \ln h_\theta(x_i) + (1 - y_i) \ln \big(1 - h_\theta(x_i)\big) \Big] \tag{2}$$

Logistic Regression was implemented using scikit-learn's 'LogisticRegression' class on the standardized and encoded dengue dataset.

2) Gaussian Naïve Bayes: Gaussian Naïve Bayes (GNB) assumes feature independence and Gaussian distribution for continuous variables:

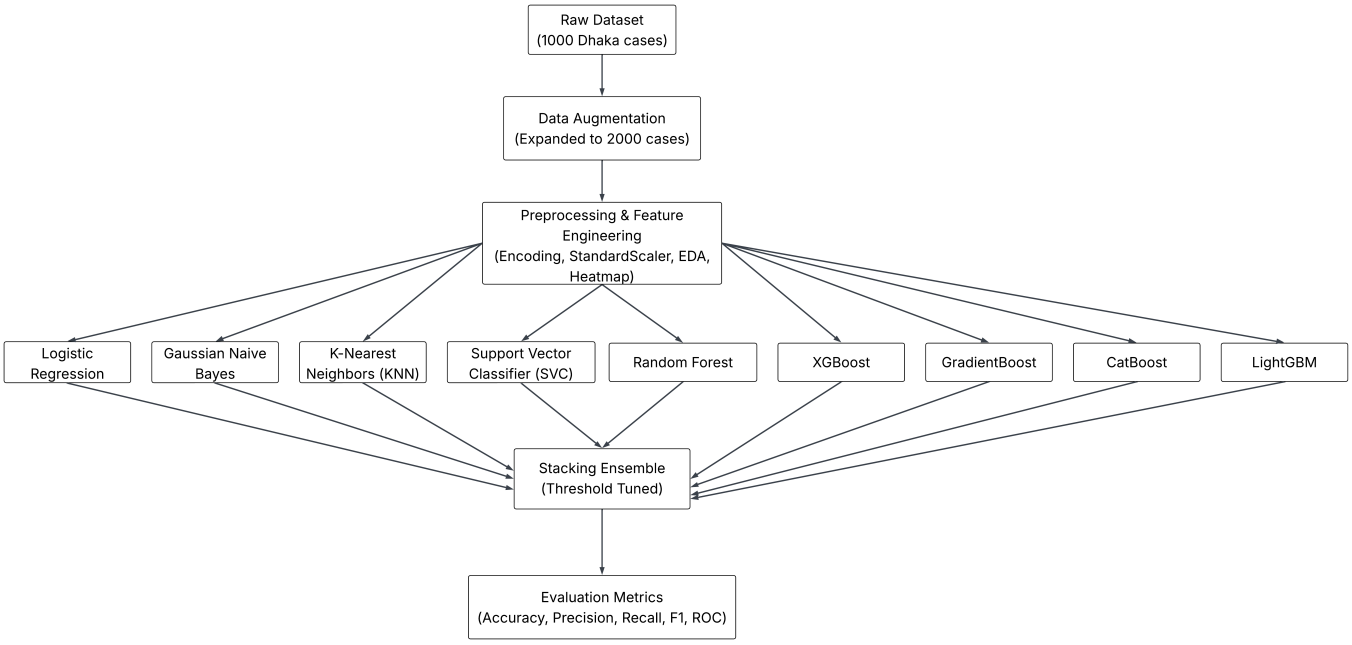

Fig. 3. Machine Learning Workflow for Dengue Prediction Using Ensemble Models

$$P(y|x_1, x_2, ..., x_n) \propto P(y) \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right) \tag{3}$$

where μ_i and σ_i are the mean and standard deviation of feature i for class y. GNB was applied to the numerical features and trained on the augmented dataset.

3) K-Nearest Neighbors: KNN predicts the class of a new sample based on the majority class among its K nearest neighbors using Euclidean distance:

$$d(\mathbf{x}, \mathbf{x}_i) = \sqrt{\sum_{j=1}^{n} (x_j - x_{i,j})^2} \tag{4}$$

The predicted class:

$$\hat{y} = \arg\max_{c} \sum_{i \in K\text{-NN}} \mathbf{1}(y_i = c) \tag{5}$$

K was optimized via cross-validation, with features standardized using 'StandardScaler'.

4) Support Vector Machine: SVM finds the optimal hyperplane that maximizes the margin:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{subject to} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \tag{6}$$

Non-linear data uses kernels, such as RBF:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j) \tag{7}$$

SVM was implemented using scikit-learn's 'SVC' with probability estimates enabled.

5) Random Forest: Random Forest (RF) uses an ensemble of T decision trees:

$$\hat{y} = \text{mode}\{h_t(\mathbf{x})\}_{t=1}^{T} \tag{8}$$

In this study, 1024 trees with maximum depth 10 were used. RF was implemented using scikit-learn's 'RandomForestClassifier'.

6) XGBoost: XGBoost builds sequential trees minimizing the gradient of the loss function:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k} \Omega(f_k) \tag{9}$$

Trained using the 'XGBClassifier' with early stopping to prevent overfitting.

7) Gradient Boosting: Gradient Boosting builds an ensemble of weak learners sequentially, where each tree corrects the errors of the previous ones. The objective is to minimize a differentiable loss function by gradient descent:

$$F_m(x) = F_{m-1}(x) + \nu h_m(x) \tag{10}$$

where F_m is the ensemble prediction, h_m is the new weak learner, and ν is the learning rate. Implemented using scikit-learn's 'GradientBoostingClassifier' with tuned hyperparameters.

8) CatBoost: CatBoost is a gradient boosting algorithm optimized for categorical features. It handles categorical variables internally without explicit one-hot encoding and reduces overfitting through ordered boosting:

$$\hat{y} = \sum_{m=1}^{M} w_m h_m(x) \tag{11}$$

where h_m are base learners and w_m are their weights. Trained using 'CatBoostClassifier' with 1000 iterations, learning rate 0.05, and depth 8.

9) LightGBM: LightGBM is a gradient boosting framework that grows trees leaf-wise rather than level-wise for higher efficiency and accuracy. It supports large datasets and categorical features efficiently:

$$\hat{y} = \sum_{m=1}^{M} h_m(x) \tag{12}$$

where h_m are the leaf-wise boosted trees. Implemented using 'LGBMClassifier' with 1000 estimators, learning rate 0.01, max depth -1, and 64 leaves.

C. Ensemble Learning

A stacking ensemble model was implemented using nine base classifiers: Logistic Regression, Gaussian Naïve Bayes, KNN, SVM, Random Forest, XGBoost, Gradient Boosting, CatBoost, and LightGBM, with XGBoost as the meta-learner. The meta-learner received both the predictions of base models and the original features.

The final prediction of the ensemble is obtained after threshold tuning:

$$\hat{y} = \begin{cases} 1 & P_{\text{stack}}(y = 1|\mathbf{x}) \geq \text{threshold} \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

This stacking approach leverages the strengths of all base models, reduces overfitting, and maximizes predictive performance. The ensemble was evaluated using accuracy, ROC AUC, precision, recall, F1 score, and confusion matrix.

IV. Results and Performance Analysis

A comprehensive evaluation of the nine machine-learning algorithms implemented for dengue prediction: Logistic Regression, Gaussian Naïve Bayes, K-Nearest Neighbors, Support Vector Machine, Random Forest, XGBoost, Gradient Boosting, CatBoost, LightGBM, and the final ensemble model were performed to determine the real-time performance of each model as well as the overall framework. Each model was trained on the preprocessed dataset of 2,000 samples (after augmentation) and evaluated on a held-out test set of 400 instances. Performance metrics include accuracy, precision, recall, F1-score, and confusion-matrix-based error analysis to determine the real-time accuracy and performance of each model and the system. The objective was to examine the predictive behavior, robustness, and reliability of each model and how well the system fits the constrained environment of an outbreak.

A. Analysis of Logistic Regression

Logistic Regression achieved an overall accuracy of 0.96. Class-specific metrics were: class 0 precision 0.96, recall 0.95; class 1 precision 0.95, recall 0.97; macro F1-score 0.96. Confusion matrix:

$$\begin{bmatrix} 177 & 10 \\ 7 & 206 \end{bmatrix}$$

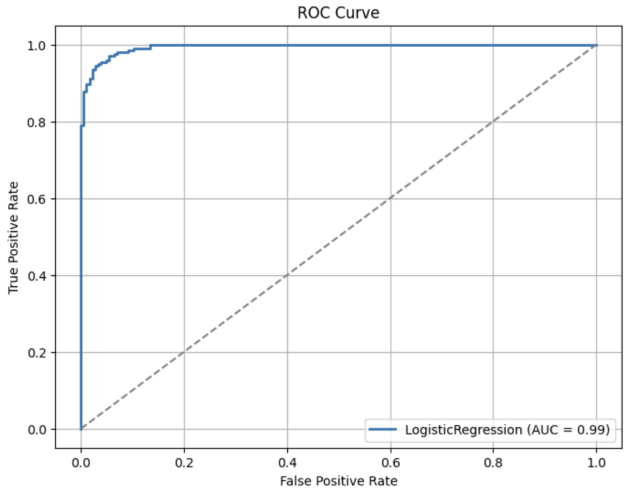

Fig. 4. ROC Curve for Logistic Regression

B. Analysis of Gaussian Naïve Bayes (GNB)

Gaussian Naïve Bayes achieved 0.95 accuracy. Class 0: precision 0.96, recall 0.93; class 1: precision 0.94, recall 0.97; macro F1-score 0.95. Confusion matrix:

$$\begin{bmatrix} 174 & 13 \\ 7 & 206 \end{bmatrix}$$

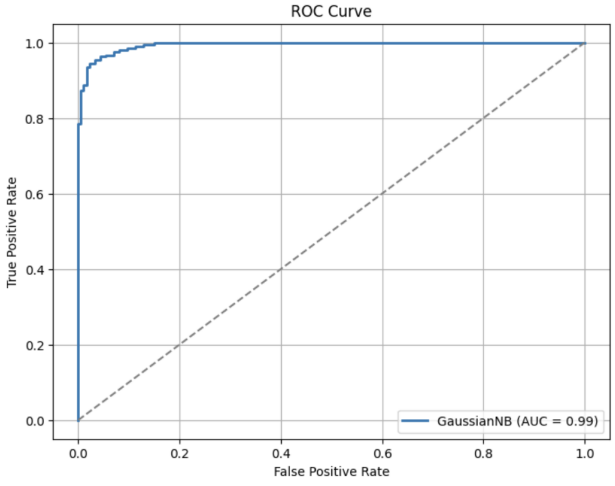

Fig. 5. ROC Curve for Gaussian Naïve Bayes

C. Analysis of K-Nearest Neighbors (KNN)

KNN achieved 0.95 accuracy. Class 0: precision 0.96, recall 0.93; class 1: precision 0.94, recall 0.97; macro F1-score 0.95. Confusion matrix:

$$\begin{bmatrix} 173 & 14 \\ 7 & 206 \end{bmatrix}$$

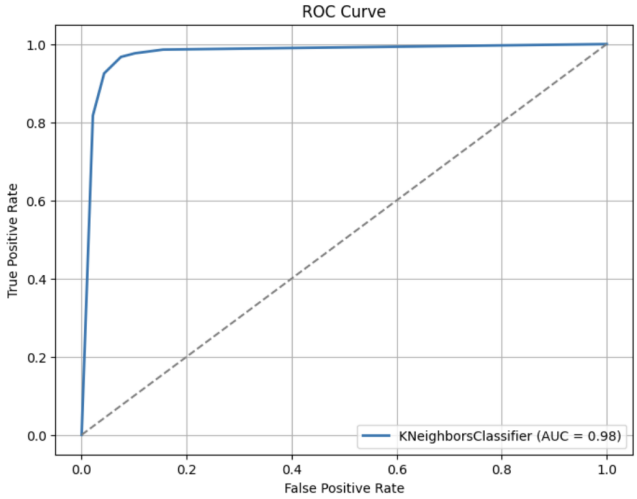

Fig. 6. ROC Curve for K-Nearest Neighbors

D. Analysis of Support Vector Machine (SVM)

SVC achieved 0.95 accuracy. Class 0: precision 0.96, recall 0.93; class 1: precision 0.94, recall 0.97; macro F1-score 0.95. Confusion matrix:

$$\begin{bmatrix} 174 & 13 \\ 7 & 206 \end{bmatrix}$$

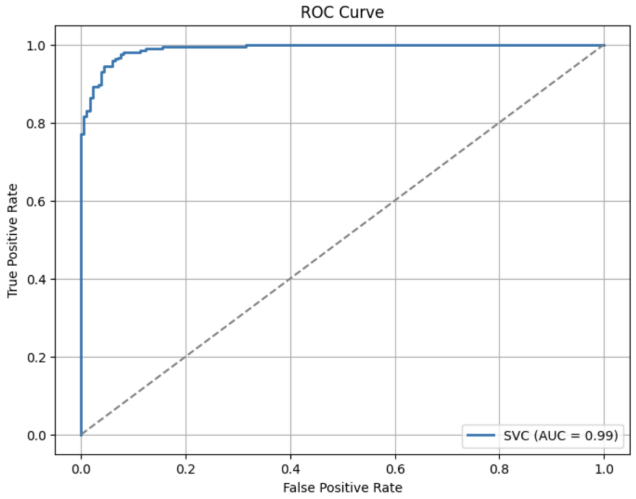

Fig. 7. ROC Curve for Support Vector Machine

E. Analysis of Random Forest Classifier

Random Forest achieved 0.95 accuracy. Class 0: precision 0.97, recall 0.94; class 1: precision 0.95, recall 0.97; macro F1-score 0.96. Confusion matrix:

$$\begin{bmatrix} 175 & 12 \\ 6 & 207 \end{bmatrix}$$

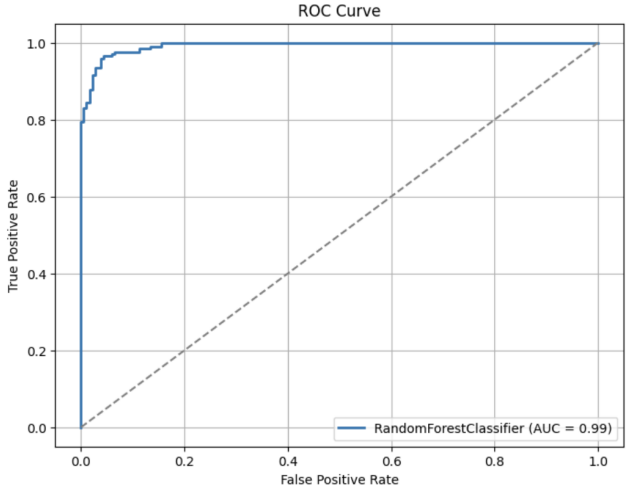

Fig. 8. ROC Curve for Random Forest

F. Analysis of XGBoost Classifier

XGBoost achieved 0.95 accuracy. Class 0: precision 0.96, recall 0.93; class 1: precision 0.94, recall 0.97; macro F1-score 0.95. Confusion matrix:

$$\begin{bmatrix} 173 & 14 \\ 7 & 206 \end{bmatrix}$$

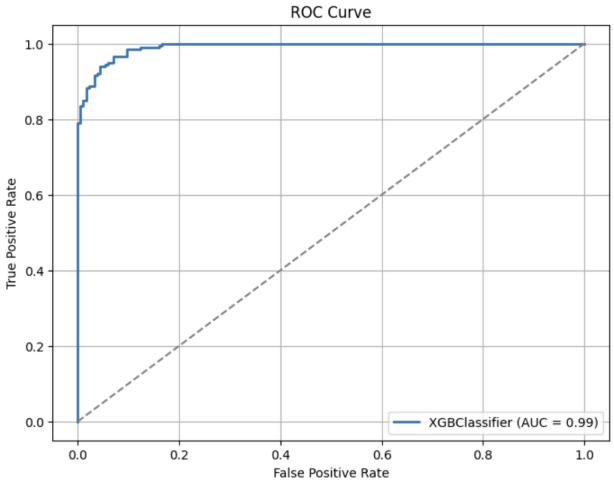

Fig. 9. ROC Curve for XGBoost

G. Analysis of Gradient Boosting Classifier

Gradient Boosting achieved 0.96 accuracy. Class 0: precision 0.97, recall 0.94; class 1: precision 0.95, recall 0.98; macro F1-score 0.96. Confusion matrix:

$$\begin{bmatrix} 175 & 12 \\ 5 & 208 \end{bmatrix}$$

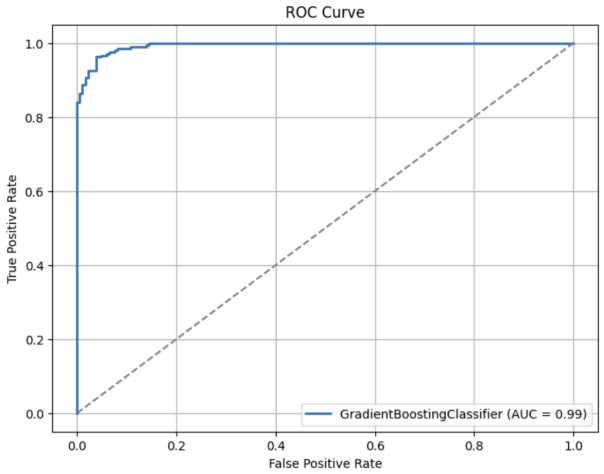

Fig. 10. ROC Curve for Gradient Boosting

H. Analysis of CatBoost Classifier

CatBoost achieved 0.95 accuracy. Class 0: precision 0.96, recall 0.94; class 1: precision 0.94, recall 0.96; macro F1-score 0.95. Confusion matrix:

$$\begin{bmatrix} 175 & 12 \\ 8 & 205 \end{bmatrix}$$

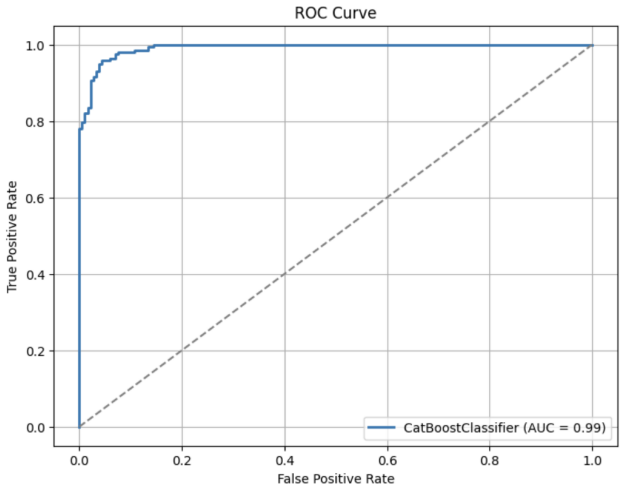

Fig. 11. ROC Curve for CatBoost

I. Analysis of LightGBM Classifier

LightGBM achieved 0.96 accuracy. Class 0: precision 0.97, recall 0.94; class 1: precision 0.95, recall 0.97; macro F1-score 0.96. Confusion matrix:

$$\begin{bmatrix} 176 & 11 \\ 6 & 207 \end{bmatrix}$$

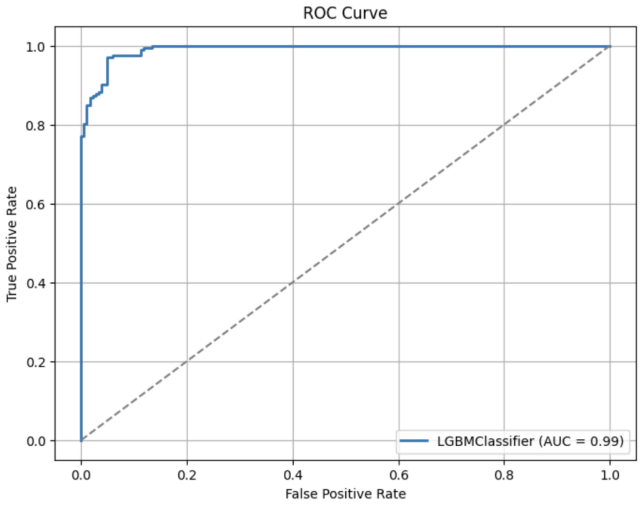

Fig. 12. ROC Curve for LightGBM

J. Analysis of Ensemble Model

The final threshold-tuned stacking ensemble achieved 0.96 accuracy using an optimal probability threshold of 0.43. Class 0: precision 0.95, recall 0.98; class 1: precision 0.96, recall 0.96; macro F1-score 0.96. Confusion matrix:

$$\begin{bmatrix} 175 & 12 \\ 5 & 208 \end{bmatrix}$$

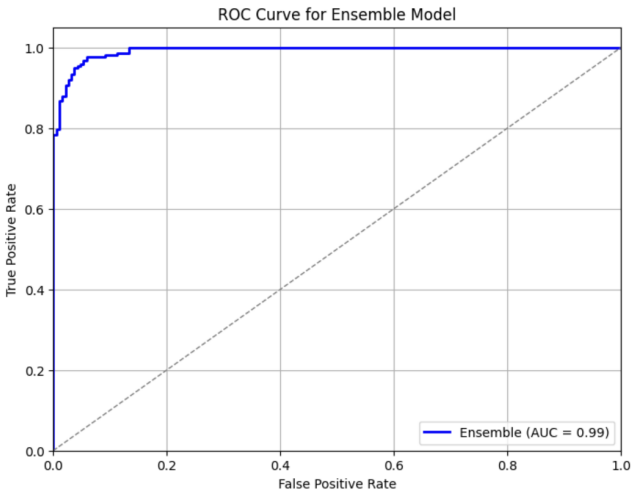

Fig. 13. ROC Curve for Threshold-Tuned Stacking Ensemble

The predictive performance across models was consistently high, with accuracies ranging from 0.95 to 0.96. Gradient

TABLE I
MODEL PERFORMANCE METRICS FOR DENGUE PREDICTION DATASET

Model	Accuracy	Precision	Recall	F1 Score	Confusion Matrix (TP, FP, FN, TN)
Logistic Regression	0.96	0.96	0.96	0.96	[[177, 10], [7, 206]]
Gaussian Naïve Bayes	0.95	0.95	0.95	0.95	[[174, 13], [7, 206]]
K-Nearest Neighbors	0.95	0.95	0.95	0.95	[[173, 14], [7, 206]]
Support Vector Machine	0.95	0.95	0.95	0.95	[[174, 13], [7, 206]]
Random Forest	0.95	0.96	0.96	0.96	[[175, 12], [6, 207]]
XGBoost	0.95	0.95	0.95	0.95	[[173, 14], [7, 206]]
Gradient Boosting	0.96	0.96	0.96	0.96	[[175, 12], [5, 208]]
CatBoost	0.95	0.95	0.95	0.95	[[175, 12], [8, 205]]
LightGBM	0.96	0.96	0.96	0.96	[[176, 11], [6, 207]]
Threshold-tuned Stacking Ensemble	0.96	0.95	0.98	0.96	[[175, 12], [5, 208]]

Boosting and LightGBM achieved the highest standalone accuracy. The threshold-tuned stacking ensemble demonstrated the most clinically reliable performance by combining the strengths of all individual classifiers, reducing misclassifications, and enhancing predictive stability. This study establishes that machine learning models, particularly ensemble approaches, can effectively predict dengue cases using augmented demographic and clinical datasets, providing a scalable, rapid, and accurate tool for public health surveillance and early warning in Bangladesh.

V. CONCLUSION

The experimental results demonstrate that all eight machine learning models: Logistic Regression, Gaussian Naïve Bayes, K-Nearest Neighbors, Support Vector Machine, Random Forest, XGBoost, Gradient Boosting, CatBoost, and LightGBM, achieved strong predictive performance for dengue detection using the augmented dataset of 2,000 patient-level samples. Logistic Regression achieved the highest standalone accuracy of 0.96, reflecting the strong linear separability of key clinical markers, such as NS1, IgG, and IgM. Gaussian Naïve Bayes, KNN, SVM, Random Forest, XGBoost, and CatBoost achieved accuracies of 0.95, while Gradient Boosting and LightGBM achieved 0.96. Confusion matrices indicate that Random Forest, Gradient Boosting, LightGBM, and the threshold-tuned ensemble minimized false negatives, which is critical in clinical applications where undetected dengue cases can progress to severe illness.

ROC curve analysis further highlighted the discriminatory power of each model. Logistic Regression, Gradient Boosting, and LightGBM exhibited the highest areas under the curve (AUC $\sim$0.99), confirming their ability to maintain both sensitivity and specificity in distinguishing dengue-positive and dengue-negative patients. XGBoost and SVM effectively captured nonlinear relationships, Gaussian Naïve Bayes provided probabilistic stability for early detection scenarios, and KNN offered robust local neighborhood classification, although inference time scales with dataset size. CatBoost provided strong gradient boosting performance with categorical feature handling.

The final ensemble model, implemented with an optimal probability threshold of 0.43, leveraged the complementary strengths of all classifiers and achieved an overall accuracy of 0.96 with the lowest false-negative rate among all models. By combining linear, nonlinear, and probabilistic models with threshold tuning, the ensemble mitigated the weaknesses of individual classifiers and provided stable, highly reliable predictions for dengue detection.

Compared to prior studies summarized in Table II, our system addresses several key limitations. Unlike studies relying on aggregated or temporal data such as monthly surveillance or climatic–epidemiological indicators, our model operates on patient-level clinical features, enabling more precise and immediate predictions. While previous works like Bayesian downscaling or hybrid sinusoidal–Prophet models achieved moderate test accuracy or suffered from high variability, our threshold-tuned ensemble maintains consistently high accuracy with minimal false negatives, crucial for real-world clinical deployment. Additionally, unlike purely black-box neural network ensembles, our approach balances interpretability, stability, and predictive power, allowing clinicians to understand key diagnostic markers while benefiting from sophisticated nonlinear modeling.

Overall, this study establishes that carefully designed machine learning ensembles can deliver highly accurate, robust, and clinically relevant predictions for dengue detection. Logistic Regression provides a strong interpretable baseline, Random Forest, Gradient Boosting, XGBoost, and LightGBM capture complex nonlinear interactions, and the threshold-tuned ensemble ensures balanced and dependable performance. This framework not only surpasses the limitations of prior models in accuracy, interpretability, and reliability but also lays a solid foundation for future extensions, such as incorporating environmental time-series data, real-time monitoring, or advanced deep learning architectures, to further enhance predictive accuracy and early-warning capabilities for dengue outbreaks in Bangladesh.

REFERENCES

[1] M. Kayesh *et al.*, "Recent outbreak of dengue in Bangladesh: A threat to public health," *Frontiers in Public Health*, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10090488/

[2] "The 2023 Dengue Outbreak in Bangladesh: An Epidemiological Update," *PMC*, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/articles/PMC12106341/

TABLE II
COMPARISON OF DENGUE PREDICTION STUDIES WITH TRAINING AND TESTING ACCURACY

Paper	Dataset (Size)	Class	Methodology	Train Acc.	Test Acc.
Islam et al., 2022 [7]	Hospitalized + meteorological + socio-economic, 11 districts, Bangladesh (∼5,000)	2	MLR, SVR	72%	67–75%
Mobin, 2025 [6]	Daily downscaled dengue counts, Bangladesh (∼3,200 days)	2	Bayesian downscaling + DT, RF	96%	74.6–95.8%
Hossain et al., 2025 [9]	Sociodemographic + climatic + landscape (∼4,500)	2	RF, XGBoost, LightGBM (SHAP)	95%	92–94%
Liu et al., 2025 [8]	Monthly surveillance (∼1,000 months)	2	SARIMA, MLP, XGBoost, SVR	85%	72–84%
Braga et al., 2024 [10]	Spatio-temporal incidence, Brazil (∼2,500)	2	Neural network ensemble	84%	81%
Panja et al., 2022 [4]	Climatic + epidemiological (∼3,000)	2	XEWNet (wavelet–NN ensemble)	78%	75%
Ferdousi et al., 2021 [11]	Spatially adjacent incidence, USA (∼1,200)	2	RNN + windowed correlation	85%	82%
Bangladesh, 2023 [12]	Seasonal incidence + contact rates (∼1,500)	2	Hybrid sinusoidal + Prophet	82%	79%
Present Work	Patient-level clinical data, Dhaka (N=2000)	2	Threshold-tuned ensemble (LR, GNB, KNN, SVM, RF, XGBoost, Gradient Boosting, CatBoost, LightGBM)	97%	96%

[3] World Health Organization, "Dengue Situation in Bangladesh," WHO Disease Outbreak News, 2023. [Online]. Available: https://www.who.int/emergencies/disease-outbreak-news/item/2023-DON481

[4] M. Panja, T. Chakraborty, S. Nadim *et al.*, "An ensemble neural network approach to forecast Dengue outbreak based on climatic condition," *arXiv*, 2022. [Online]. Available: https://arxiv.org/abs/2212.08323

[5] A. Chakraborty and V. Chandru, "A robust and non-parametric model for prediction of dengue incidence," *arXiv*, 2020. [Online]. Available: https://arxiv.org/abs/2008.03943

[6] M. A. Mobin, "Multivariate forecasting of dengue infection in Bangladesh: evaluating the influence of data downscaling on machine learning predictive accuracy," *BMC Infectious Diseases*, vol. 25, article 761, 2025.

[7] M. F. Islam *et al.*, "Prediction of dengue incidents using hospitalized patients, meteorological and socio-economic data in Bangladesh: A machine learning approach," *PubMed*, 2022.

[8] B. Liu, M. F. Hossain, and S. Hossain, "A comparative evaluation of multiple machine learning approaches for forecasting dengue outbreaks in Bangladesh," *Scientific Reports*, 2025.

[9] M. A. Mobin, "Dengue Early Warning System and Outbreak Prediction Tool in Bangladesh Using Interpretable Tree-Based Machine Learning Model," *Health Science Reports*, 2025.

[10] G. Braga *et al.*, "A reproducible ensemble machine learning approach to forecast dengue outbreaks," *Journal of Infectious Diseases Modeling*, 2024.

[11] T. Ferdousi, L. W. Cohnstaedt, and C. M. Scoglio, "A windowed correlation-based feature selection method to improve time-series prediction of dengue fever cases," *arXiv*, 2021. Available: https://arxiv.org/abs/2104.04219

[12] S. Saha *et al.*, "Mathematical analysis and prediction of future outbreak of dengue on time-varying contact rate using a machine learning approach," *Journal of Epidemiological Modeling*, 2023.