Bangladesh Army University of Science and Technology (BAUST), Saidpur

Department of Computer Science and Engineering (CSE)
Course Title: Machine Learning Sessional
Course Code: CSE 4140

Dengue Prediction Using Ensemble Machine Learning Models

Presented By:

MST. Sumya Jafrin ID: 220201013

A. K. M. Masudur Rahman ID: 220201065

Presented To:

Engr. Rohul Amin Designation: Lecturer

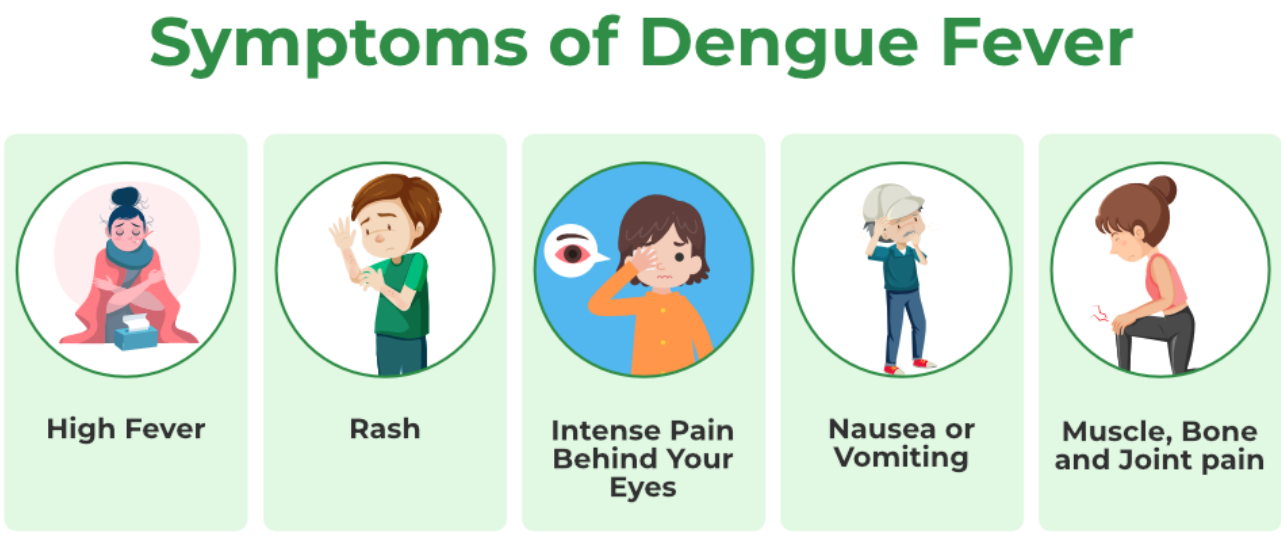
Nadim Reza Designation: Lecturer

Introduction

❑ New System for Early Dengue Prediction

Our system uses machine-learning models to quickly predict dengue infection using clinical, demographic, and environmental data.

- **Detects and Predicts:** Early dengue infection.
- **Goal:** Improve public health response during dengue outbreaks.



Objectives

To predict dengue infection early using machine-learning models.

To reduce diagnostic delays during peak outbreak seasons.

To provide reliable decision support for healthcare professionals.

Related Works

In [7], Islam et al., in 2022, proposed a predictive framework using hospitalized patient data combined with meteorological and socio-economic features across 11 districts in Bangladesh.

Shortcomings:

- Relied on only two models (MLR and SVR), limiting capture of non-linear relationships.
- Focused on aggregated predictions rather than patient-level outcomes.
- Limited applicability for early detection in clinical settings.

Related Works

In [6], Mobin, in 2025, proposed a multivariate forecasting approach using Bayesian downscaling to convert monthly dengue counts into daily estimates.

Shortcomings:

- Focused mainly on population-level temporal trends.
- Not optimized for real-time clinical decision support.
- Did not incorporate serological or patient-specific demographic features.

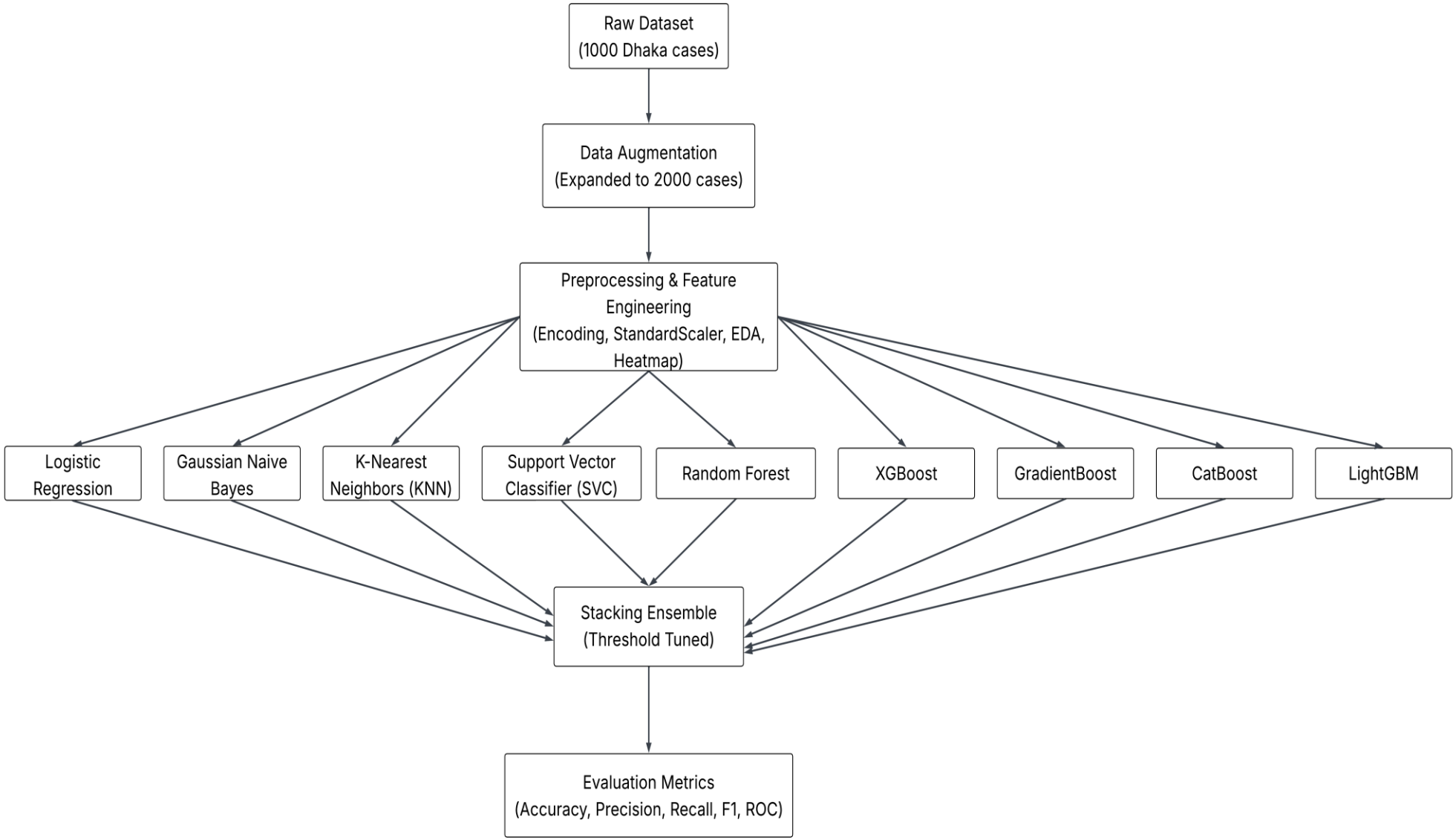
Related Works

In [9], Hossain et al., in 2025, implemented Random Forest, XGBoost, and LightGBM models using sociodemographic, climatic, and landscape features with SHAP-based interpretability.

Shortcomings:

- Focused on aggregated population risk, not patient-specific predictions.
- Relied heavily on environmental variables, limiting early clinical diagnosis.
- Did not fully leverage serological markers for precise predictions.

Outline of Methodology



Dataset

- The initial Dataset contained 1000 patient data collected from Dhaka.
- Total of 9 features.
- 2 classes: 0 (Dengue-Negative) and 1 (Dengue-Positive)
- **Features:** NS1, IgG, IgM, Age, Gender, AreaType, HouseType, District, etc.

	A	B	C	D	E	F	G	H	I	J
1	Gender	Age	NS1	IgG	IgM	Area	AreaType	HouseType	District	Outcome
2	Female	45	0	0	0	Mirpur	Undevelop	Building	Dhaka	0
3	Male	17	0	0	1	Chawkbaz	Developed	Building	Dhaka	0
4	Female	29	0	0	0	Paltan	Undevelop	Other	Dhaka	0
5	Female	63	1	1	0	Motijheel	Developed	Other	Dhaka	1
6	Male	22	0	0	0	Gendaria	Undevelop	Building	Dhaka	0
7	Female	36	0	0	1	Dhanmonc	Developed	Other	Dhaka	0
8	Female	15	0	0	1	New Marke	Undevelop	Building	Dhaka	0
9	Male	26	0	0	0	New Marke	Developed	Other	Dhaka	0
10	Female	31	0	0	1	Dhanmonc	Undevelop	Tinshed	Dhaka	0
11	Female	10	0	0	1	Sher-e-Bar	Developed	Tinshed	Dhaka	0
12	Female	31	1	1	0	Kafrul	Undevelop	Building	Dhaka	1
13	Male	10	0	0	0	Dhanmonc	Developed	Tinshed	Dhaka	0
14	Female	13	1	1	0	Pallabi	Undevelop	Building	Dhaka	1
15	Female	43	1	1	0	Mohamma	Developed	Building	Dhaka	1

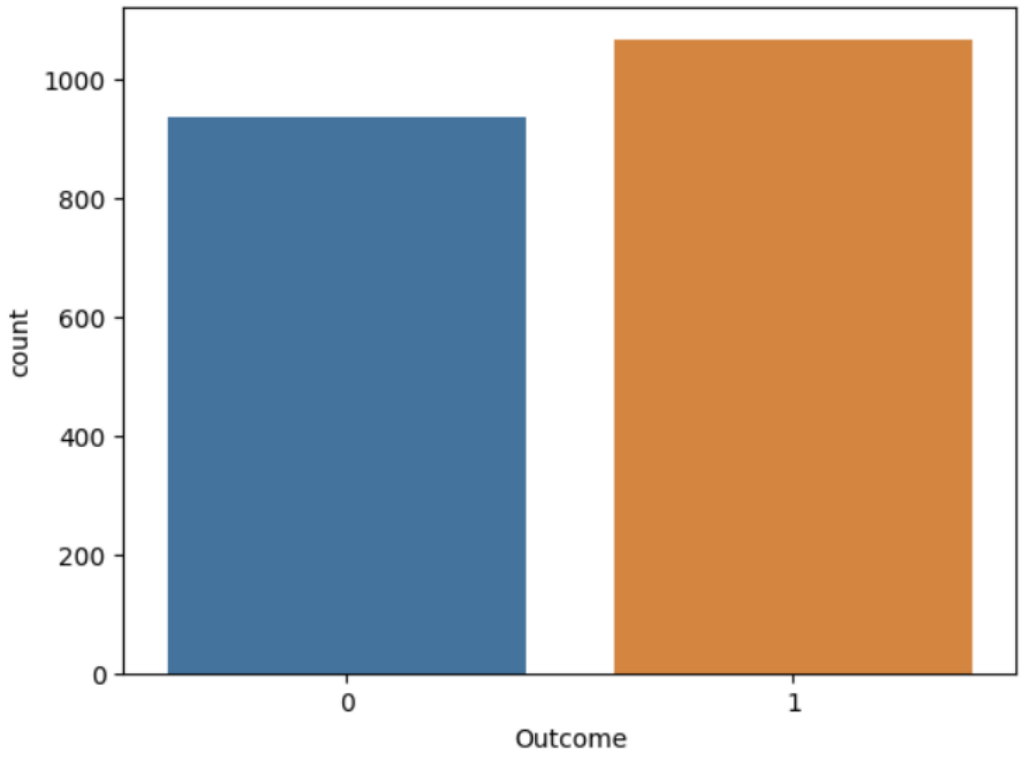
Data Augmentation

- To prevent the model overfitting, Data Augmentation was performed.
- Doubled the Dataset to 2000 data.

Techniques applied:

- **SMOTE (Synthetic Oversampling)** to balance dengue-positive & negative classes
- **Random Feature Perturbation**: small, plausible variations to numeric clinical values (NS1, IgG, IgM)
- **Categorical Resampling**: proportionally varied features like AreaType and HouseType

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

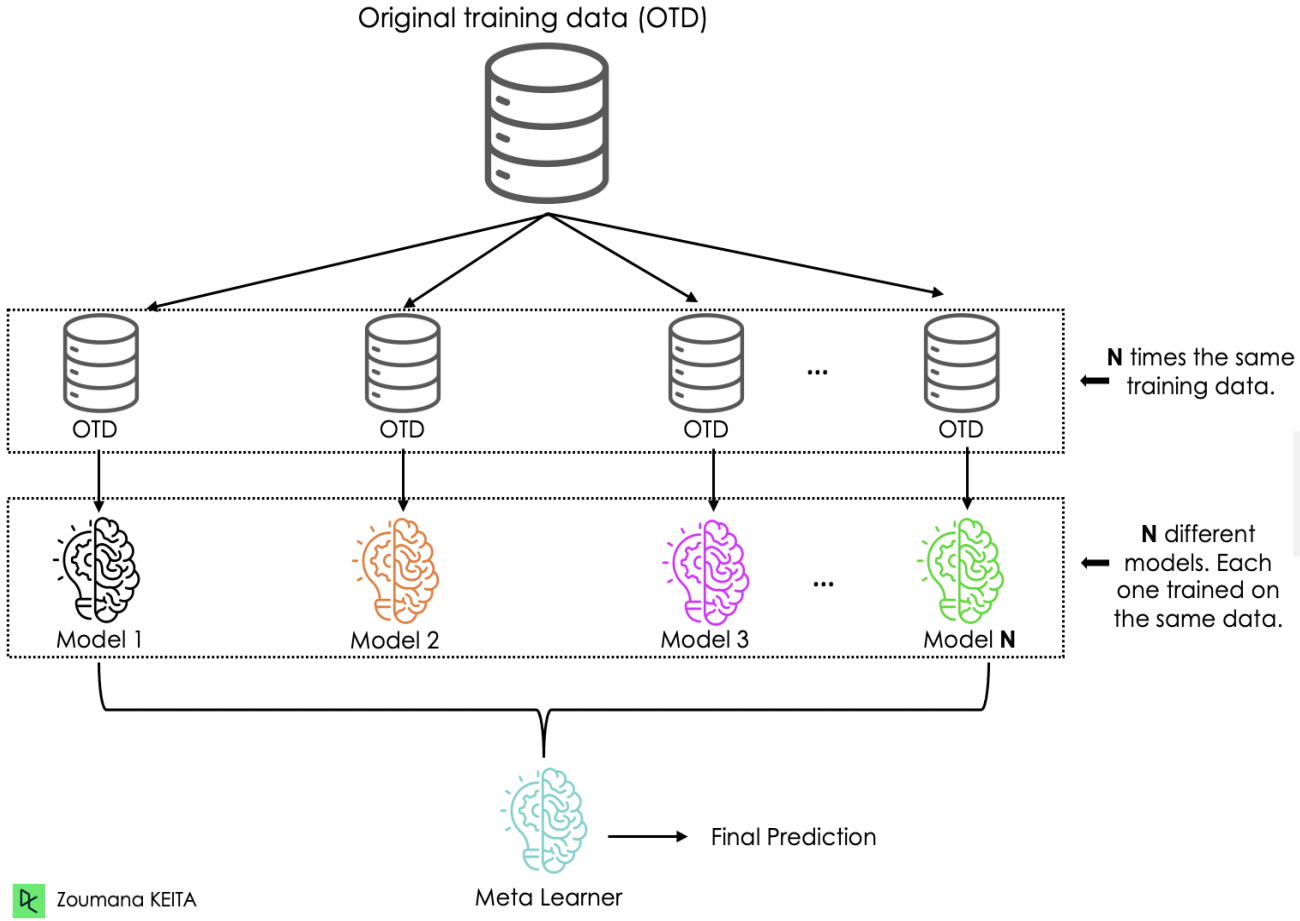


Model Implementations

We implemented 9 machine learning models as Base Learners:

1. Logistic Regression (LR) – Linear baseline
2. Gaussian Naïve Bayes (GNB) – Probabilistic
3. K-Nearest Neighbors (KNN) – Local neighborhood
4. Support Vector Machine (SVM) – Kernel-based non-linear
5. Random Forest (RF) – Bagging ensemble of trees
6. XGBoost – Gradient boosting
7. Gradient Boosting (GB) – Tree-based boosting
8. CatBoost – Handles categorical features efficiently
9. LightGBM – Leaf-wise gradient boosting for speed

Stacking Ensemble



Stacking

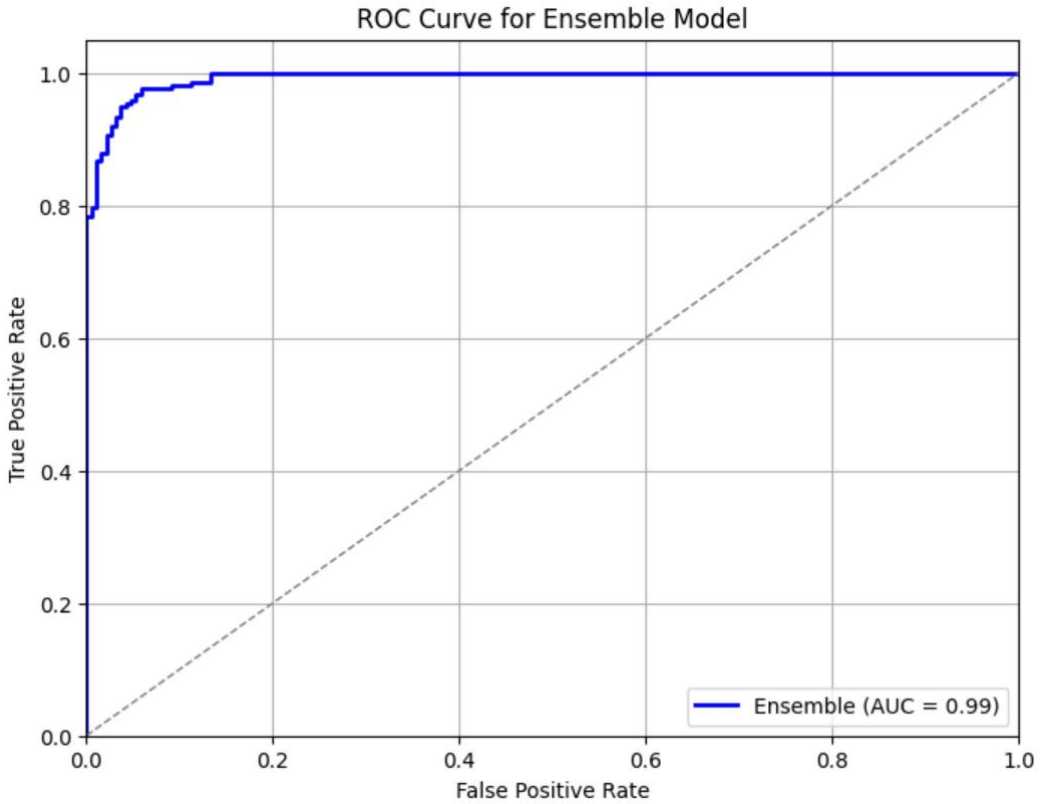
We implemented a Threshold-Tuned Stacking Ensemble Model for overall performance and accuracy.

Implementation:

- Base models: LR, GNB, KNN, SVM, RF, XGBoost, GB, CatBoost, LightGBM
- Meta-learner: XGBoost
- Threshold tuning: Optimal threshold = 0.43
- Goal: Combine base learners to improve predictive stability and minimize false negatives

Outcomes

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]]**



Optimal threshold: 0.43, Best Accuracy: 0.9600

Comparison with Previous Work

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%
Braga et al., 2024 [10]	Spatio-temporal incidence, Brazil (~2,500)	2	Ensemble neural network combining spatial & temporal predictors	84%	81%
Present Work	**Patient-level clinical data, Dhaka (N=2,000)**	**2**	**Threshold-tuned ensemble (LR, GNB, KNN, SVM, RF, XGBoost, Gradient Boosting, CatBoost, LightGBM)**	**97%**	**96%**

Limitations of Our Work

Dataset is relatively small and collected from limited regions.

Seasonal & environmental patterns are **not captured**, reducing predictive power for outbreak forecasting.

Threshold Tuning Depends on Current Dataset.

No Real-Time Deployment or API Integration

Scope for Future Work

Expand	Expand Dataset Size & Regional Coverage
Add	Add Symptom-Level, CBC Test Data, Environmental & Seasonal Features.
Use	Use Deep Learning Models (e.g. TabNet, Neural networks with attention, Hybrid ANN + boosting models etc.)
Convert	Convert Model to a Deployed System.

References

[6] Mobin, A. (2025). Bayesian downscaling for daily dengue forecasting in Bangladesh.

[7] Islam, M., et al. (2022). Dengue prediction using hospitalized patient data and meteorological features in Bangladesh.

[9] Hossain, S., et al. (2025). Random Forest, XGBoost, and LightGBM for dengue prediction with SHAP interpretability.

[10] Braga, et al. (2024). Ensemble neural network combining spatial and temporal predictors for dengue incidence forecasting in Brazil.

THANK YOU!

ANY QUESTIONS?