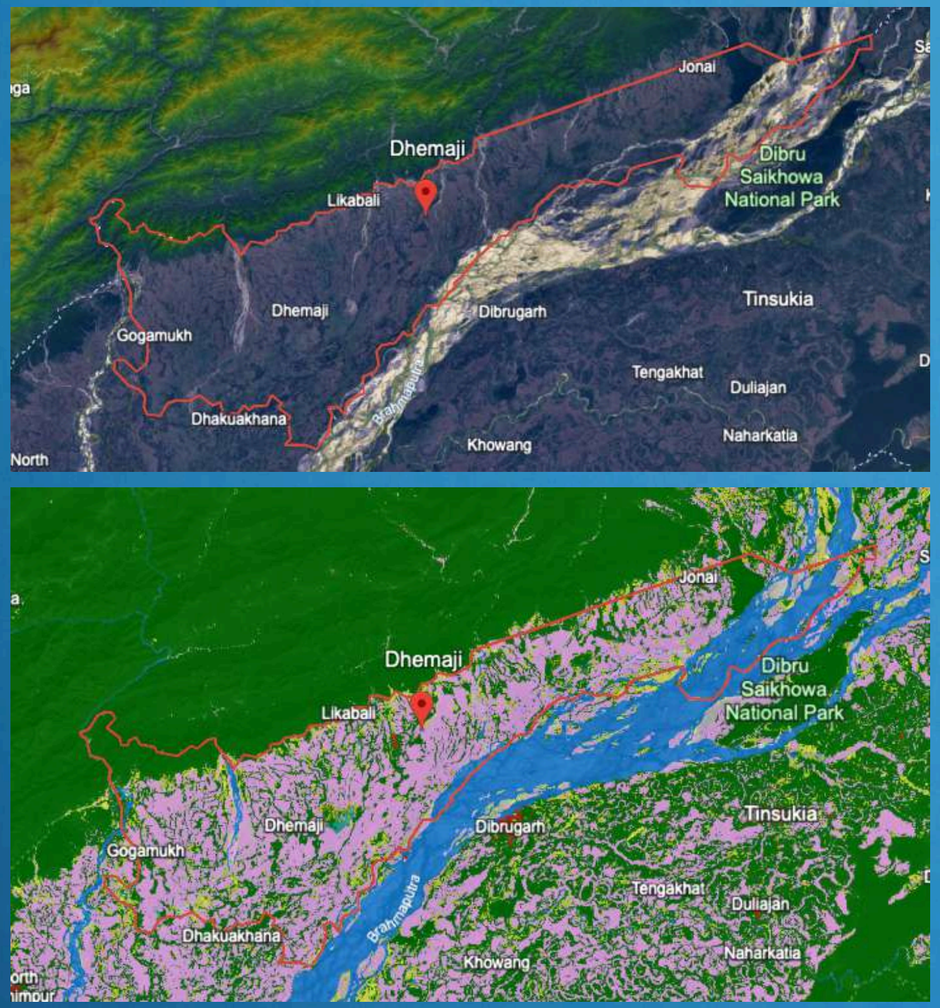
Flood Prediction in the Brahmaputra Basin

Using Terrain Modeling and Weather Conditions



Avi Gautam,
Rakshaan Thareja,
Shourrya Gupta

Machine Learning & Pattern Recognition

Prof. Siddharth

Problem Statement

~25,000 Ha of Crops destroyed in 2019

₹1700+ Cr Annual economic damage in Assam

Per-farm annual flood losses around ₹36,000–₹39,000

Why flood prediction matter ?

Floods are among the most destructive natural disasters in India, causing large-scale economic damage, agricultural loss, and population displacement every monsoon season.

Why Existing Systems Struggle ?

Traditional rainfall-based warning systems often fail in Brahmaputra floodplains due to:

- cloud cover during monsoon,
- lack of real-time river gauge coverage,
- poor temporal flood labeling,
- and weak modeling of upstream hydrology.

Why Dhemaji?

- Located within the Brahmaputra floodplain
- Experiences recurrent seasonal flooding and river erosion
- Highly agriculture-dependent and flood-vulnerable
- Suitable for studying spatiotemporal flood dynamics

Literature Review

PAPER	STUDY REGION	DATASET	MODELS	BEST ROC-AUC	KEY LIMITATION
Debnath et al. (2023)	Brahmaputra Basin (Assam, NE India)	1,000 inv. points DEM, NDVI, LULC 18 conditioning vars	RF, SVM TOPSIS, VIKOR	> 0.70	Static susceptibility only; no temporal forecasting; no SAR dynamic labeling
Saravanan et al. (2023)	Idukki District (Kerala, India)	Sentinel-1 SAR inventory 16 GIS + hydrological vars Kerala flood dataset	AdaBoost, GB XGBoost, CatBoost SGB	0.92	Static susceptibility mapping; no temporal forecasting; no date-wise flood prediction
Halder et al. (2024)	Regional multi-factor (GEE-derived inventory)	Sentinel-1 via GEE 18 geo-env. + hydro. vars Multi-factor flood dataset	LR, SVM, RF XGBoost, DNN Stacking Ensemble	0.965	Static spatial mapping; no temporal forecasting across seasons; no date-wise prediction
Feizbahr et al. (2025)	River basin hazard zones (multi-source RS)	Sentinel-1, Sentinel-2 SRTM DEM 14 conditioning factors	RF, XGBoost CNN (deep learning) SHAP (XAI)	~0.98 (CNN)	Static spatial susceptibility; no temporal flood evolution; no strict holdout year test

The Research Gap

No Temporal Dimension

Every prior study produces a static susceptibility map only a snapshot.None forecast when flooding will occur. Seasonal dynamics and monsoon progression are completely ignored.

No SAR-Dynamic Labels

Ground truth is fixed from a single SAR acquisition or small inventory. No study uses multi-date SAR labels that track actual flood extents across 150+ events.

No Strict Holdout Validation

Models are validated on randomly split data from the same time period , optimistic and prone to leakage. None reserve an entire unseen monsoon year for testing.

Dynamic SAR Labeling

150 unique flood acquisition dates across 6 monsoon seasons. Multi-date SAR labels capture spatial flood extent variation for each date.

Temporal Forecasting

1.25M spatio-temporal samples with cumulative antecedent rainfall (Rain_3day, Rain_5day) to model soil saturation dynamics day by day.

Unseen-Year Holdout

Entire 2024 monsoon season reserved as the holdout set , never seen during training. Strict temporal generalization for real-world deployment.

Brahmaputra Focus

Applied to Dhemaji, Assam the most flood-vulnerable district in India. Region-specific features including NDVI, elevation, river proximity.

Debnath et al. (2023)

Debnath, J., Sahariah, D., Mazumdar, M., Lahon, D., Meraj, G., Hashimoto, S., Kumar, P., Singh, S. K., Kanga, S., Chand, K., & Saikia, A. (2023). Evaluating Flood Susceptibility in the Brahmaputra River Basin: An Insight into Asia's Eastern Himalayan Floodplains Using Machine Learning and Multi-Criteria Decision-Making. Earth Systems and Environment, 7(4), 733-760. https://doi.org/10.1007/s41748-023-00358-w

Most geographically **relevant study** to our work because it studies:

- Brahmaputra Basin
- Assam floodplains
- Dhemaji-like terrain conditions

Dataset:

- 1000 inventory points
- DEM, rainfall, NDVI, LULC, river distance
- 18 flood-conditioning variables

Models: RF, SV, TOPSIS ,VIKOR

Results:

- ROC-AUC > 0.70 on all models and highest 0.96 RF
- RF performed best
- Elevation & river proximity dominant

Limitation:

- Static susceptibility only
- No temporal forecasting
- No SAR dynamic labeling

Our Improvement:

- Dynamic SAR labels
- Temporal flood prediction
- 1.25M data rows
- Unseen monsoon year testing

Saravanan et al. (2023)

Saravanan, S., Abijith, D., Reddy, N. M., KSS, P., Janardhanam, N., Sathiyamurthi, S., & Sivakumar, V. (2023). Flood susceptibility mapping using machine learning boosting algorithms techniques in Idukki district of Kerala, India. Urban Climate, 49, 101503.

Dataset:

- Sentinel-1 SAR flood inventory
- 16 GIS + hydrological variables
- Kerala flood susceptibility dataset

Models:

- AdaBoost
- Gradient Boosting
- XGBoost
- CatBoost
- SGB

Results:

- Best ROC-AUC = 0.92
- Boosting models performed strongest
- SAR + ML highly effective

Limitation:

- Static susceptibility mapping
- No temporal forecasting
- No date-wise flood prediction

Our Improvement:

- Dynamic SAR labels
- 150 flood dates
- Temporal flood forecasting
- 1.25M spatio-temporal samples

Halder et al. (2024)

Halder, A., et al. (2024). SAR-driven flood inventory and multi-factor ensemble machine learning for flood susceptibility mapping. International Journal of Digital Earth. Taylor & Francis.

Dataset:

- Sentinel-1 SAR flood inventory via Google Earth Engine (GEE)
- 18 geo-environmental + hydrological variables
- Regional multi-factor flood susceptibility dataset

Models:

- Logistic Regression (LR)
- Support Vector Machines (SVM) & Random Forest (RF)
- XGBoost & Deep Neural Networks (DNN)
- Stacking Ensemble Model (Meta-Classifier)

Results:

- Best ROC-AUC = 0.965 (achieved by Stacking Ensemble)
- Advanced ensemble frameworks vastly outclassed standalone models
- SAR-derived inventory proved highly reliable in cloudy monsoon regions

Limitation:

- Static spatial susceptibility mapping
- No temporal forecasting components across different seasons
- No date-wise flood prediction mechanics

Our Improvement:

- Dynamic multi-date SAR labels mapping explicit spatial variations
- 150 total flood dates tracking 6 consecutive monsoon seasons
- True temporal flood forecasting driven by cumulative antecedent rainfall features
- 1.25M spatio-temporal samples evaluated on an unseen holdout year

Feizbahr et al. (2025)

Feizbahr, M., et al. (2025). Flood susceptibility mapping using machine learning and remote sensing approaches. Remote Sensing, 17(1), Article 112. MDPI. Susceptibility Mapping Using Machine Learning and Geospatial-Sentinel-1 SAR Integration for Enhanced Early Warning Systems

Dataset:

- Multi-source remote sensing database (Sentinel-1 SAR, Sentinel-2 Optical, and SRTM DEM)
- 14 geo-environmental conditioning factors spanning topography, multi-scale hydrology, and vegetation indices
- Highly robust spatial flood inventory mapping operational river basin hazards

Models:

- Random Forest (RF)
- eXtreme Gradient Boosting (XGBoost)
- Convolutional Neural Networks (CNN-based deep learning models)
- SHAP (SHapley Additive exPlanations) for Explainable AI (XAI)

Results:

- Best ROC-AUC = ~0.97 → XG
- XGBoost (0.97) and RF (0.96) demonstrated excellent performance for tabular processing
- SHAP interpretation successfully prioritized elevation and hydrological proximity as main drivers

Limitation:

- Primarily focused on static spatial susceptibility zonation
- Lacks temporal flood evolution or continuous day-wise tracking across seasons
- Evaluation did not implement a strict, forward-looking year-based holdout test

Our Improvement:

- Explicit spatio-temporal flood prediction instead of static hazard mapping
- Dynamic multi-date target labels synchronized over 150 unique SAR acquisition timelines
- Integration of hydrology-aware lagged rainfall features (Rain_3day and Rain_5day) to capture soil saturation dynamics
- Rigorous validation strategy using an entirely unseen monsoon year (2024 holdout) to ensure strict temporal generalization

Data Preprocessing

Datasets:

→ Rainfall

CHIRPS (Climate Hazards Group InfraRed Precipitation with Station data)

35+ years
5km resolution

→ Elevation

NASA SRTM DEM: provides 30m elevation data used to derive terrain height and slope.

→ Surface runoff

ERA5-Land: provides daily runoff data representing hydrological flow conditions contributing to flooding.

→ Tree Cover

MODIS VCF Tree Cover Dataset: provides vegetation/tree cover density information affecting runoff and infiltration.

→ Dynamic Flood Labels

Sentinel-1 SAR: Floods were identified by detecting changes in water signatures over time, allowing date-wise flood mapping.

→ Distance to River

HydroSHEDS dataset
Convert river vectors into spatial geometry
↓
For each grid cell: find nearest distance to river

Feature Preprocessing:

Raster-to-Grid Numerical Extraction:

Raster Datasets
(DEM, Rainfall, Tree Cover, Water)

↓

Extract Pixel Values + Coordinates
(latitude, longitude)

↓

Convert to Numerical Grid Data

↓

Spatially Align All Datasets
onto Common 500m Grid

↓

Generate Unified ML Feature Table

Derived Features:

DEM + Rainfall Time-Series

↓

Derive Terrain Features
(slope calculated from elevation gradients)

↓

Generate Rolling Rainfall Features
(rain_3day = cumulative rainfall over previous 3 days,
rain_5day = cumulative rainfall over previous 5 days)

↓

Compute Rainfall Anomaly
(current rainfall - seasonal average rainfall)

↓

Capture Runoff, Soil Saturation
& Extreme Rainfall Conditions

↓

Create Enhanced ML Features
for Flood Prediction

Feature Preprocessing:

Class Imbalance Handling :

Highly **Imbalanced** Dataset
(94.3% non-flood, 5.7% flood)

↓

Apply **Higher Weight** to **Flood** Samples
(flood rows weighted 8.8×)

↓

Train ML Model

↓

Lower Decision Threshold
(0.5 → 0.4)

↓

Improve Flood Recall
(0.821 → 0.854)

↓

Reduce Missed Flood Events

Distance to Major River :

HydroSHEDS River Network

↓

Filter Major Rivers

↓

Compute Nearest Distance
for Each 500m Grid Cell to a river

↓

Generate **dist_to_major_river** Feature

↓

Used in ML Flood Prediction

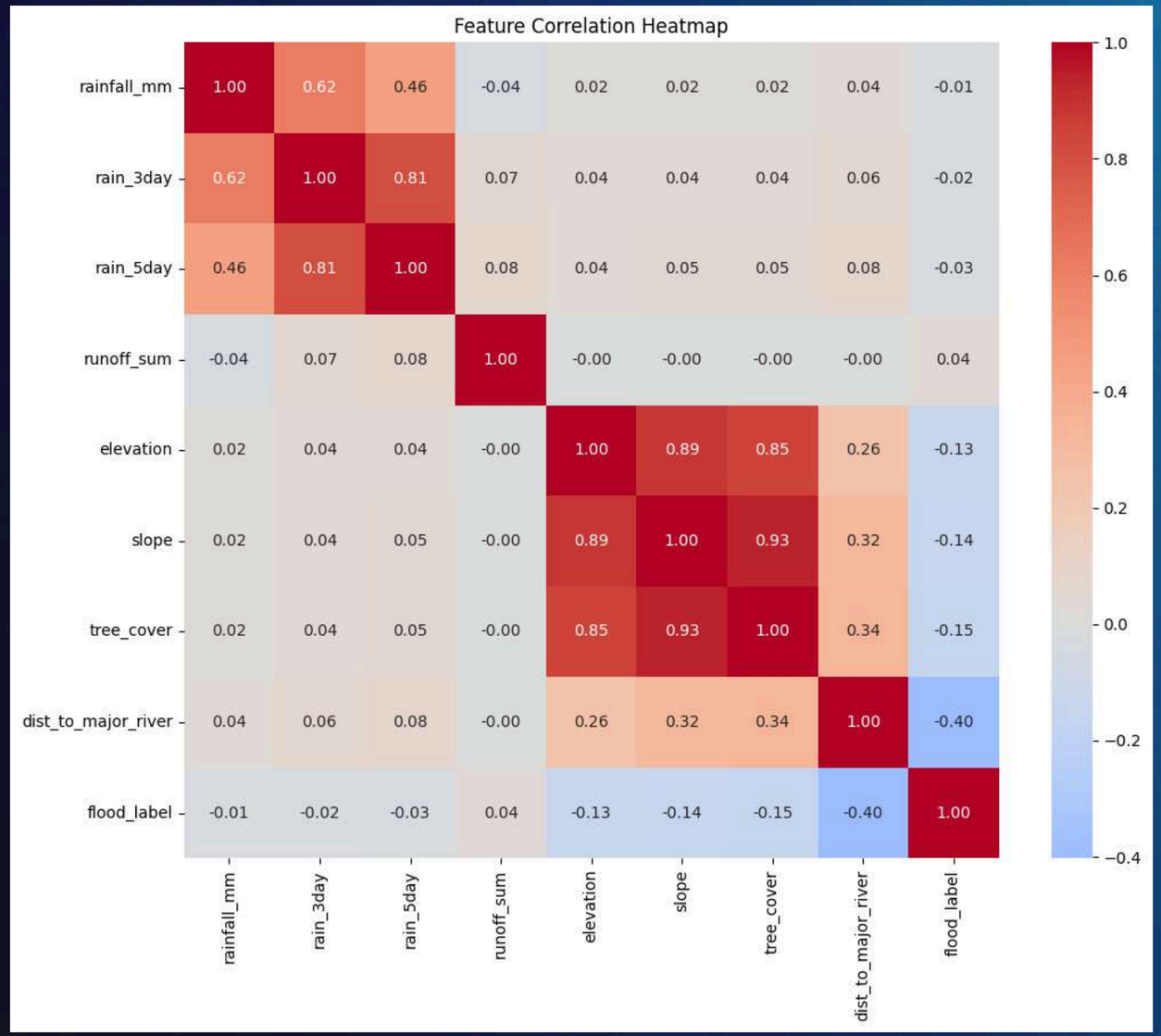
Cells closest to the river flooded 368x more than the farthest cells, making this the strongest predictor at 46% feature importance.

Combined Datasets

~1.25 million spatio-temporal rows and 8,658 spatial grid cells
2019–2024 monsoon seasons where Sentinel-1 satellite images were collected on 150 different days/timestamps across the study period.

latitude	longitude	date	flood_label	year	rain_3day	rain_5day	rain_anomaly	runoff_anomaly	rainfall_mm	elevation	slope	tree_cover	dist_to_major_river
27.40086195	94.35230008	03-06-2019	0	2019	0	0	-15.0263664	-0.006221371	0	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	15-06-2019	0	2019	56.56185	56.56185	41.5354836	-0.002644848	56.56185	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	27-06-2019	0	2019	56.56185	56.56185	-15.0263664	0.012661187	0	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	09-07-2019	0	2019	98.955634	98.955634	27.3674176	0.028192377	42.393784	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	21-07-2019	0	2019	146.628414	203.190264	89.2082636	0.000834174	104.23463	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	23-07-2019	0	2019	146.628414	203.190264	-15.0263664	0.00240865	0	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	02-08-2019	0	2019	104.23463	146.628414	-15.0263664	-0.001512004	0	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	14-08-2019	0	2019	28.828966	175.45738	13.8025996	-0.000392859	28.828966	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	16-08-2019	0	2019	28.828966	133.063596	-15.0263664	-0.003475768	0	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	26-08-2019	0	2019	28.828966	28.828966	-15.0263664	-0.005541926	0	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	28-08-2019	0	2019	15.723216	44.552182	0.696849601	-0.005750735	15.723216	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	07-09-2019	0	2019	34.305303	63.134269	3.555720601	-0.004537248	18.582087	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	09-09-2019	0	2019	90.051557	90.051557	40.7198876	0.00314827	55.746254	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	19-09-2019	0	2019	74.328341	90.051557	-15.0263664	-0.001816311	0	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	21-09-2019	0	2019	55.746254	90.051557	-15.0263664	-0.00304119	0	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	09-06-2020	0	2020	0	74.328341	-15.0263664	-0.00381245	0	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	21-06-2020	0	2020	44.8005	100.546754	29.7741336	-0.004858386	44.8005	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	03-07-2020	0	2020	44.8005	44.8005	-15.0263664	-0.003164001	0	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	05-07-2020	0	2020	112.24145	112.24145	52.4145836	-0.002485708	67.44095	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	15-07-2020	0	2020	67.44095	112.24145	-15.0263664	-0.00021632	0	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	17-07-2020	0	2020	82.413885	127.214385	-0.053431399	-0.000168515	14.972935	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	23-07-2020	0	2020	105.402055	172.843005	75.4027536	0.00170204	90.42912	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	27-07-2020	0	2020	105.402055	172.843005	-15.0263664	-0.002610967	0	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	29-07-2020	0	2020	90.42912	105.402055	-15.0263664	0.004036711	0	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	08-08-2020	0	2020	30.678972	136.081027	15.6526056	-0.003193439	30.678972	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	10-08-2020	0	2020	30.678972	121.108092	-15.0263664	-0.001854129	0	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	20-08-2020	0	2020	74.195078	74.195078	28.4897396	0.0013382	43.516106	94.87132353	1.53030646	15.25	31265.02971
27.40086195	94.35230008	22-08-2020	0	2020	43.516106	74.195078	-15.0263664	-0.002700819	0	94.87132353	1.53030646	15.25	31265.02971

Correlation Matrix



Key Insights

- Main Driver: **dist_to_major_river** correlates strongest (0.40); closer means higher flood probability.
- Rainfall Insufficiency: Local rainfall has near-zero correlation (−0.01 to −0.03); local storms do not drive Brahmaputra floods.
- Runoff Signal: runoff_sum captures a more meaningful physical signal than raw rainfall with a correlation of 0.04.

METHODOLOGY

Flood Susceptibility Modeling Framework

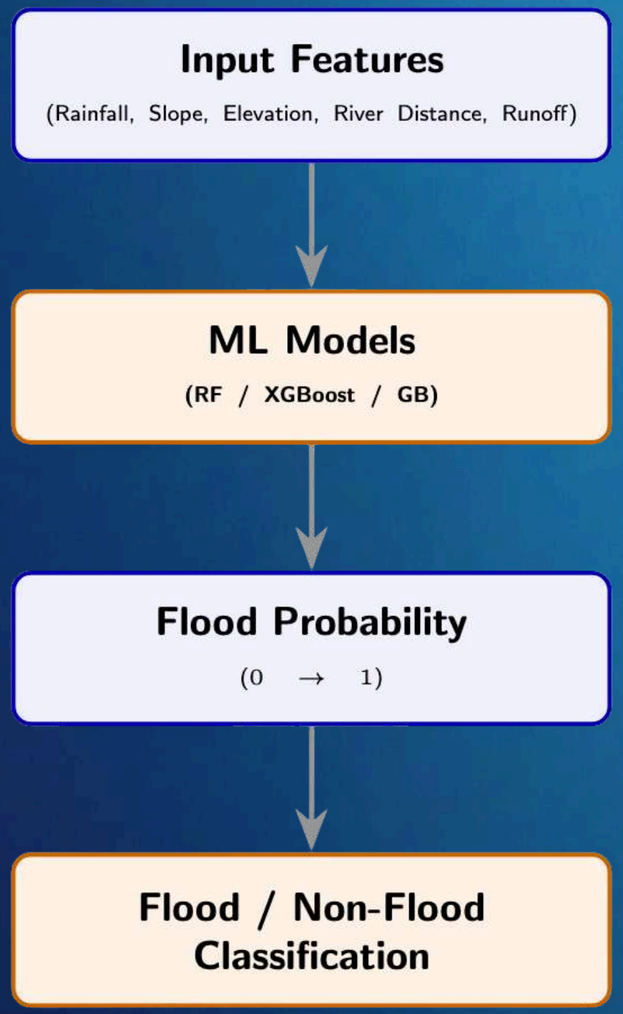
Objective:

Predict flood occurrence for each 500m × 500m grid cell on a given date

Binary Classification:

1 = Flooded

0 = Non-Flooded



Training Set

2019–2023

1,090,005 rows

Testing Set

2024 Monsoon

165,337 rows

Machine Learning Models Used

We ran 10 models and below are the metrics calculated for each model

Model	Precision	Recall	F1	ROC_AUC	Rank
Gradient Boosting (Tuned)	0.825	0.854	0.839	0.99	1
Random Forest	0.734	0.936	0.823	0.992	2
Neural Network	0.817	0.807	0.812	0.988	3
Extra Trees	0.679	0.95	0.792	0.991	4
XGBoost	0.676	0.945	0.788	0.991	5
LightGBM	0.667	0.954	0.785	0.992	6
AdaBoost	0.815	0.744	0.778	0.981	7
CatBoost	0.648	0.96	0.774	0.99	8
KNN	0.683	0.531	0.598	0.914	9
Logistic Regression	0.286	0.967	0.441	0.972	10

Final Model

Evaluated across 11 models, Gradient Boosting achieved the best balance of precision and F1 score:

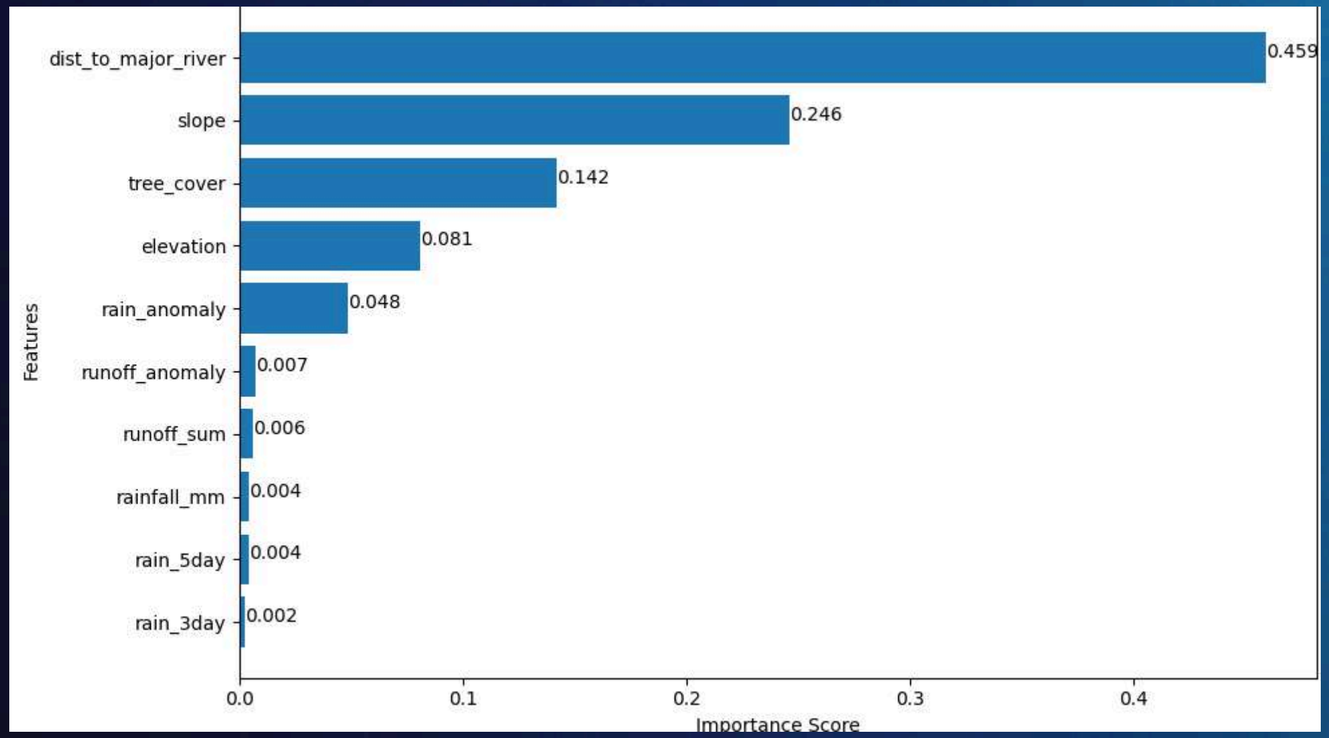
Gradient Boosting = Precision 0.845 Recall 0.821 F1 0.839

Hyperparameter Tuning:

Parameter Values	Tested	Final Value
n_estimators	200 , 300, 500	200
max_depth	3, 5, 7	5
Learning Rates	0.05, 0.1	0.1
Sub Sample	0.6, 0.8, 1	0.8
Min_Samples_leaf	10, 20, 50	20
Max_features	3, 5	3

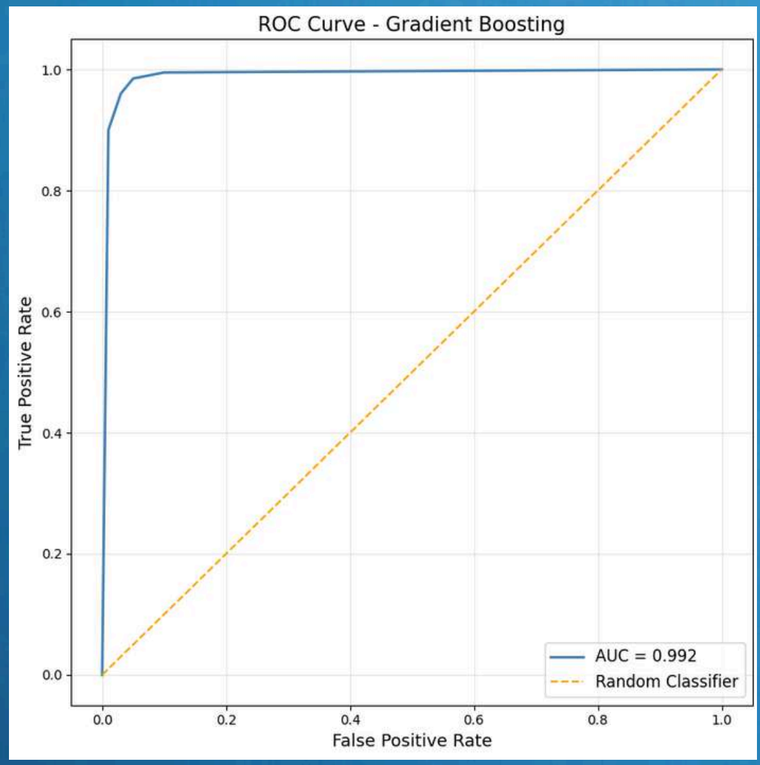
Model Evaluation

Feature Importance



Spatial cross-validation (trained on 80% of grid cells) and evaluates whether the model can predict floods in completely unseen geographic regions, and the minimal AUC drop (0.990 to 0.982) demonstrates spatial **generalization** rather than location **memorization**.

ROC Curve



The high AUC reflects the physical reality of Brahmaputra flooding the same low-lying riverside areas flood repeatedly every monsoon season, creating strong and consistent spatial patterns. Cells closest to the river flood 368x more than the farthest cells, making the class separation predictable.

MODEL BENCHMARKING COMPARISONS

iterative internal model benchmarking:

Version	Key_Change	Precision	Recall	F1	ROC_AUC	Rows	Dates	Years
V1 Original	Static extent labels	0.56	0.84	0.67	0.95	100000	13	1
V2 SAR + Distance	Dynamic SAR labels + river	0.71	0.92	0.80	0.976	100000	13	1
V3 Upstream Rain	Added upstream rainfall	0.70	0.92	0.80	0.976	100000	13	1
V4 Multiyear	Extended to 6 monsoon seasons	0.77	0.91	0.84	0.992	1090005	150	6
V5 Reduced Features	10 clean features + runoff	0.74	0.94	0.82	0.992	1090005	150	6
FINAL GB Tuned	Gradient Boosting + threshold	0.825	0.854	0.839	0.99	1090005	150	6

- Existing literature shows ensemble tree-based models such as Random Forest, XGBoost, and Stacking Ensembles consistently outperform traditional statistical methods for flood susceptibility mapping, achieving ROC-AUC values between ~0.92–0.97.

Limitations:

1. Incomplete Sentinel-1 spatial coverage
2. No upstream river discharge data
3. Possible spatial autocorrelation - leading to an inflated ROC curve
4. Local rainfall may not capture basin-scale hydrology
5. No real-time deployment pipeline yet

Scope for Future Work:

1. Integrate upstream river discharge and basin-scale flow data.
2. Include soil moisture and groundwater saturation variables.
3. Model flood propagation dynamics along the Brahmaputra floodplain.
4. Develop near real-time flood prediction pipeline using live satellite feeds.

References:

1. Debnath, P., Das, P., & Roy, S. (2023). Evaluating flood susceptibility in the Brahmaputra River Basin using machine learning and multi-criteria decision-making approaches. Springer.
2. Saravanan, S., Abijith, D., Reddy, N. M., KSS, P., Janardhanam, N., Sathiyamurthi, S., & Sivakumar, V. (2023). Flood susceptibility mapping using machine learning boosting algorithms techniques in Idukki district of Kerala, India. Urban Climate, 49, 101503.
3. Halder, A., et al. (2024). SAR-driven flood inventory and multi-factor ensemble machine learning for flood susceptibility mapping. International Journal of Digital Earth. Taylor & Francis.
4. Feizbahr, M., et al. (2025). Flood susceptibility mapping using machine learning and remote sensing approaches. Remote Sensing, 17(1), Article 112. MDPI.

Thank You