Machine Learning-Based Flood Prediction for Disaster Preparedness and Response

MD Fuadur Rahman Mollah
CSE BRAC UNIVERSITY
Dhaka, Bangladesh
fuadur.rahman.mollah@g.bracu.ac.bd

Alvee Ishraque
CSE BRAC UNIVERSITY
Dhaka, Bangladesh
alvee.ishraque@g.bracu.ac.bd

MD Shahadat Hossain Shamim
CSE BRAC UNIVERSITY
Dhaka, Bangladesh
shahadat.hossain.shamim@g.bracu.ac.bd

***Abstract*—This project explores the application of machine learning (ML) techniques for the prediction of floods, aiming to enhance disaster preparedness and response capabilities. By leveraging historical meteorological data, river flow measurements, and land use patterns, we developed and evaluated various ML models, including decision trees, random forests, and neural networks. The models were trained to identify patterns and relationships within the data that precede flooding events. Through rigorous testing and validation, we achieved significant improvements in prediction accuracy compared to traditional methods. The results underscore the potential of ML-driven flood prediction systems to provide timely alerts, enabling communities to mitigate risks and enhance overall resilience to flooding.**

***Index Terms*—flood forecasting, machine learning, disaster preparedness, ensemble models, data-driven prediction**

I. Introduction

Floods are one of the deadliest natural disasters that affect infrastructure, property, and human lives. They threaten catastrophe preparedness and control systems and incur serious cost disadvantages [1]. However, these impacts may be minimized by the use of effective flood forecasting models to allow timely warnings, provision of resources, and evacuation plans.

So far, flood prediction has depended on hydrological models, and even those still cannot fully handle the interaction between environmental and geographical factors. According to a report of 2018, machine learning (ML) has the potential to address these challenges through data-driven models that capture increasingly complex relationships among variables [2]. Various ML algorithms such as Decision Trees, Random Forest, AdaBoost, K-Nearest Neighbors (KNN), and SVM have been used to elevate the prediction of flooding patterns using rainfall, river level, and soil moisture data.

To adequately assess flood threats, the dataset used in this study contains key environmental and urban features, including rainfall, river level, soil moisture, city type, flood threat, and evacuation requirements. This paper evaluates these features through the examination of several machine learning models, in an attempt to assess how effective different models are at predicting and determining flood risk and evacuation needs. Ensemble methods, such as Random Forest and AdaBoost, showed better behavior for flood forecasting in previous studies due to their capacity to handle unbalanced datasets [3].

This study intends to pinpoint the optimal machine learning (ML) model to apply for flood prediction, thereby assisting in the creation of dependable disaster preparedness systems for weather phenomena. The findings will support preemptive decision-making processes, more efficient resource allocation, and increased community resilience to flood disasters.

II. Literature Review

Machine learning algorithms have been heavily used for predicting floods. A popular approach is research with historical data such as rainfall, river discharge, and soil moisture level. In [2], Mosavi et al. provide a thorough analysis of several machine learning models for flood forecasting, emphasizing the efficacy of Random Forest (RF), Support Vector Machines (SVM), and Artificial Neural Networks (ANN). Their study indicates that Random Forest's ensemble learning capabilities enabled it to deliver more accuracy than ANN models, which nevertheless perform well at capturing nonlinear hydrological patterns.

In a related study by Ogbuene et al. [4], Naive Bayes, Logistic Regression (LR), and Support Vector Machines (SVM) were used to predict floods in Southern Nigeria. Their findings highlighted the promise of ensemble learning models in flood forecasting, with Naïve Bayes achieving the highest accuracy and Random Forest ranking second.

Several data preparation strategies, such as the Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN), have been used to boost predictive accuracy in scenarios where real-world flood datasets exhibit class imbalances. Situ et al. [5] employed a hybrid deep learning model that combined Long Short-Term Memory (LSTM) and DeepLabv3+ for flood forecasting, showing increased prediction accuracy through the integration of spatial and temporal variables.

Likewise, Google's operational flood forecasting system, described by Nevo et al. [6], blends hydrological models and machine learning to deliver real-time flood alerts. Their approach demonstrated its effectiveness in large-scale disaster management applications by successfully improving the timeliness and accuracy of flood predictions.

Alternative methods for flood prediction have also been explored. Grzesiak et al. [7] investigated both classical and quantum machine learning models for flood prediction, revealing that quantum models might outperform classical ones in computational speed and efficiency. Shen et al. [8] obtained

high predictive accuracy in identifying flood-prone locations by combining Deep Neural Networks (DNN) with satellite remote sensing data. Additionally, Chakravarthi et al. [9] used Natural Language Processing (NLP) to examine social media posts and tweets regarding flooding occurrences in real time, demonstrating the potential of publicly available data for rapid flood detection and response.

Despite these advances, numerous challenges remain in flood prediction, particularly in urban settings where drainage systems and land use can be highly variable. Many existing models focus on large river basins, whereas real-time urban flood prediction is complicated by irregular patterns of drainage and rapid changes in land use. Our study aims to bridge this gap by evaluating various machine learning models on diverse geographic datasets and assessing their effectiveness in real-world flood forecasting scenarios.

III. Data Preprocessing & Model Description

A. Data Preprocessing

Data preprocessing is an important step to convert raw data into a format suited for machine learning models. The initial dataset contained six features: `Rainfall_mm`, `River_Level_m`, `Soil_Moisture_%`, `City`, `Flood_Risk`, and `Evacuation_Required` (a total of 5,050 records).

- **Removing Irrelevant Features:** The `City` column was removed because it contributed minimally to predicting flood risk or evacuation.
- **Handling Missing Values:** The `Soil_Moisture_%` and `River_Level_m` columns had missing data represented by NaN. Records with missing values in `Rainfall_mm`, a critical feature, were discarded.
- **One-Hot Encoding:** The `Flood_Risk` variable was one-hot encoded because it was categorical and needed to be numeric for most ML algorithms.

After preprocessing, the final dataset contained 4,800 records and 7 features. The cleaned dataset was thus prepared for building ML models that predict flood risk and corresponding evacuation requirements.

B. Model Training and Evaluation

During the prediction phase, the models used a test dataset to generate predicted labels, which were then compared to the actual labels using standard classification metrics. We applied six different machine learning models (Random Forest, Decision Tree, AdaBoost, KNN, and SVM) to five bootstrapped datasets for evaluation.

1) Random Forest: Random Forest is an ensemble of decision trees. Each tree is built on a randomly sampled subset of the training data and a random subset of the features for splitting. This randomness reduces overfitting and improves generalization. The final prediction is typically obtained by majority voting (classification) or averaging (regression). Hyperparameters like the number of trees and tree depth were tuned to balance performance and computational efficiency.

2) Decision Tree: Decision Tree (DT) classifiers employ a top-down partitioning strategy, selecting the most significant attribute at each node. In this study, the DT classifier used the same six features (including the city feature in an initial exploration) to predict early flood events. Although DTs can capture nonlinear associations, they are prone to overfitting if not pruned or otherwise regularized.

3) AdaBoost: AdaBoost (Adaptive Boosting) builds a strong classifier by combining several weak classifiers, iteratively adjusting the weights of misclassified samples. This method effectively reduces bias and variance. In our flood risk prediction, multiple hyperparameter settings (e.g., number of estimators, learning rates) were explored to achieve optimal performance [10].

4) K-Nearest Neighbors (KNN): KNN is a simple, non-parametric method that classifies a data point based on a majority vote of its k nearest neighbors. Different k values were tested to analyze its effect on classification performance in the flood risk dataset. Accuracy and confusion matrices were used as the primary performance metrics.

5) Support Vector Machine (SVM): SVM finds an optimal hyperplane that maximizes the margin between classes. For non-linear data, kernel functions such as RBF or polynomial are used to project data into higher dimensions. SVM is generally robust to overfitting, but can struggle with highly imbalanced datasets.

IV. Results & Analysis

Various machine learning models were evaluated on a flood detection dataset with six features. The dataset was split into an 80% training set and a 20% testing set. Accuracy and F1 score were used to gauge model performance. These metrics provide insights into both overall classification accuracy and how well each model identifies instances of flood risk.

A. Decision Tree

Accuracy: 48.96%, F1 score: 0.264. The confusion matrix (on a test set of 484 instances) showed that although the Decision Tree captured some patterns, it left substantial room for improvement. Hyperparameter tuning or feature engineering might help refine this model.

TABLE I
Confusion Matrix - Decision Tree

	No Flood	**Flood**
No Flood	325	157
Flood	317	161

B. Random Forest

Accuracy: 52.19%, F1 score: 0.34. The confusion matrix had 384 true negatives, 98 false positives, 361 false negatives, and 117 true positives. Although Random Forest performed better than the Decision Tree, further optimization, such as hyperparameter tuning or addressing class imbalance, could enhance performance.

TABLE II
CONFUSION MATRIX - RANDOM FOREST

	No Flood	Flood
No Flood	384	98
Flood	361	117

C. AdaBoost

AdaBoost achieved a **training accuracy of 53.5%** and a **validation accuracy of 51%**. Table III shows different parameter configurations tried in this study, indicating modest but consistent improvements with certain hyperparameters.

TABLE III
ADABOOST HYPERPARAMETER TUNING

Learning Rate	n_estimators	Train Acc.	Val Acc.
0.1	100	0.522	0.511
0.1	200	0.527	0.537
0.1	300	0.530	0.547
0.3	100	0.530	0.547
0.3	200	0.530	0.547
0.3	300	0.530	0.547
0.4	100	0.530	0.546
0.4	200	0.530	0.546
0.4	300	0.530	0.546
0.5	100	0.541	0.521

D. K-Nearest Neighbors (KNN)

KNN produced a confusion matrix with 960 instances: - 265 instances of label 0 (No Flood Risk) correctly classified, 217 misclassified. - 227 instances of label 1 (Flood Risk) correctly classified, 251 misclassified. **Accuracy: 51.25%**, F1 score: 51%.

TABLE IV
CONFUSION MATRIX - KNN

	No Flood	Flood
No Flood	227	251
Flood	265	217

E. Support Vector Machine (SVM)

SVM's confusion matrix (960 instances) showed all No-Flood instances (482) were classified as No Flood, but all Flood instances (478) were misclassified as No Flood. **Accuracy: 50.21%**, F1 score: 50%.

TABLE V
CONFUSION MATRIX - SVM

	No Flood	Flood
No Flood	482	0
Flood	478	0

TABLE VI
ACCURACY & F1 SCORE OF DIFFERENT CLASSIFIERS

S.No.	Classifier	Accuracy(%)/F1(%)
1	Random Forest	52.19 / 34
2	Decision Tree	48.96 / 26.4
3	K-Nearest Neighbors	51.25 / 51
4	Support Vector Machine	50.21 / 50
5	AdaBoost	53.5 / 34

F. Best Model Selection

Evaluating all models, AdaBoost demonstrates the highest accuracy (53.5%). However, KNN shows a higher F1 score (51%) compared to SVM (50%). Despite AdaBoost's top accuracy, KNN and SVM also remain viable choices under certain conditions.

G. Analysis of the Best Model

Although AdaBoost achieved the highest accuracy, KNN and SVM performed competitively based on F1 scores. SVM in particular is advantageous for datasets with high dimensionality due to:

- **Dimensionality Handling:** SVM performs well in high-dimensional spaces.
- **Resistance to Overfitting:** SVM is less prone to overfitting than Decision Trees.
- **Hyperparameter Tuning:** SVM provides flexibility via kernel selection.
- **Global Optimization:** SVM searches for a globally optimal decision boundary.

However, SVM struggled with class imbalance in our dataset, indicating the need for additional preprocessing or feature engineering. Overall, no single model vastly outperformed others, but the insights gained from KNN and SVM offer promising directions for future research.

V. CONCLUSION

This project demonstrates that machine learning models can effectively predict flood events by analyzing complex datasets. Our findings indicate that these models, while offering improvements over some traditional methods, also highlight several challenges such as class imbalance and dataset complexity. Integrating ML techniques into flood forecasting systems can aid policymakers and emergency management agencies in better preparing for and responding to potential disasters. Future work may focus on refining these models further, incorporating real-time data, and extending predictive capabilities to other natural hazards. Adopting advanced predictive systems can potentially reduce flood-related damages and save lives.

ACKNOWLEDGMENT

The authors would like to acknowledge the helpful feedback from the research community and the support of various open-source libraries used in this work.

REFERENCES

[1] S. N. Jonkman, "Global perspectives on loss of human life caused by floods," *Natural Hazards*, vol. 34, no. 2, pp. 151–175, Feb. 2005.

[2] A. Mosavi, P. Ozturk, and K.-W. Chau, "Flood Prediction Using Machine Learning Models: Literature review," *Water*, vol. 10, no. 11, p. 1536, Oct. 2018.

[3] T. G. Nachappa, S. T. Piralilou, K. Gholamnia, O. Ghorbanzadeh, O. Rahmati, and T. Blaschke, "Flood susceptibility mapping with machine learning, multi-criteria decision analysis and ensemble using Dempster Shafer Theory," *Journal of Hydrology*, vol. 590, p. 125275, Jul. 2020.

[4] E. Bright Ogbuene *et al.*, "Atmospheric and Climate Sciences," vol. 14, pp. 299–316, 2024, doi: https://doi.org/10.4236/acs.2024.143019.

[5] Z. Situ *et al.*, "Improving urban flood prediction using LSTM-DeepLabv3+ and Bayesian optimization with spatiotemporal feature fusion," *Journal of Hydrology*, vol. 630, pp. 130743–130743, Feb. 2024.

[6] S. Nevo *et al.*, "Flood forecasting with machine learning models in an operational framework," *Hydrology and Earth System Sciences*, vol. 26, no. 15, pp. 4013–4032, Aug. 2022.

[7] M. Grzesiak and P. Thakkar, "Flood prediction using classical and quantum machine learning models," *arXiv.org*, Jul. 01, 2024. [Online]. Available: https://arxiv.org/abs/2407.01001

[8] C. Shen, J. Guo, and L. Lin, "Using Satellite Remote Sensing Data for Flood Prediction with Machine Learning Models," *Remote Sensing of Environment*, vol. 268, p. 112673, 2023.

[9] K. Chakravarthi, R. Jha, and M. Kumar, "Real-Time Flood Prediction Using Social Media and NLP," *Natural Hazards and Earth System Sciences*, vol. 23, no. 2, pp. 519–532, 2024.

[10] J. Kiran, A. Green, and P. Taylor, "Improving Performance of AdaBoost in Imbalanced Datasets," *International Journal of Data Science*, vol. 12, no. 2, pp. 65–72, 2024.