2025-11

Rainfall trends and flood prediction in Chattogram district: a data-driven approach

Chowdhury, Pratya

School of Business and Entrepreneurship, Independent University, Bangladesh (IUB)

Rainfall Trends and Flood Prediction in Chattogram District: A Data-Driven Approach

Pratya Chowdhury[1], Ohcitya Bhattacharjee[2]
[1]Institute of Forestry and Environmental Science, University of Chittagong, Chittagong, Bangladesh
[2]Department of Computer Science & Engineering, Chittagong University of Engineering and Technology
chowdhurypratya0066@gmail.com[1], *bhattacharjeeohcitya@gmail.com*[2]

PAPER ID : ICEBTM-25-1169

Abstract – **Flooding poses severe socio-economic challenges in Bangladesh, with nearly one-third of the country highly vulnerable. Despite its recurring impact, flood research remains limited. This study addresses the gap by analyzing rainfall trends and developing a data-driven flood prediction framework for the Chattogram district, a region prone to extreme weather. Analysis of historical data revealed intensifying monsoon rainfall, with critical flood thresholds identified: annual rainfall above 3000 mm and June–September totals exceeding 2000 mm. Several machine learning models were applied to forecast flood events. Among them, the Random Forest (RF) model demonstrated the highest performance, achieving 92% accuracy and a 93% F1-score. Logistic Regression performed poorly due to its limitations with complex patterns, while K-Nearest Neighbors and Decision Trees showed moderate accuracy. Gradient Boosting was consistent but outperformed by RF. These results highlight the effectiveness of combining rainfall thresholds with machine learning to enhance early warning systems and flood risk management.**

Keywords – **Chattogram, flood prediction, machine learning, rainfall trends, Random Forest**

I. INTRODUCTION

Chattogram district, in southeastern Bangladesh, is an important economic and environmental region vulnerable to flooding due to its diverse terrain, including hills, coastal plains, and river networks. The tropical monsoon climate brings heavy rainfall between June and September, often causing riverine floods, flash floods, and urban waterlogging. Urbanization and unplanned land use have worsened flood risks by reducing natural drainage and increasing surface runoff, leading to floods and landslides, especially in hilly areas. Notable events include the devastating 1991 cyclone and 2017 monsoon floods, which caused significant damage and displacement. Climate change intensifies these challenges by altering rainfall patterns and raising sea levels. Rapid urban growth further strains drainage infrastructure, resulting in frequent waterlogging. Flooding impacts public health, agriculture, and the economy, with losses reaching billions annually. This study analyzes rainfall trends and employs data-driven models to improve flood prediction and risk management, aiming to support climate resilience and disaster preparedness in the region.

II. METHODOLOGY

To develop an accurate flood prediction system for Chattogram, a structured data-driven methodology was followed, incorporating environmental data processing and machine learning techniques. Data Collection: Historical rainfall and flood data were gathered from the Bangladesh Agricultural Research Council (BARC) and the Bangladesh Meteorological Department (BMD). These included monthly and annual rainfall figures along with flood occurrence indicators.

a. *Data Preprocessing:* Data preprocessing is an essential step in the data analysis pipeline that involves transforming raw data into a A flow diagram of the proposed algorithm's data processing phases is shown in Fig 1
b. *Preliminary Data Analysis*: Data analysis is a crucial step in the preprocessing phase, ensuring that the data is clean, well-understood, and ready for modeling. This step ultimately leads to more accurate and reliable machine learning models. During this phase, the entire dataset was examined, and issues such as empty rows and inconsistencies in column naming—some columns had names in uppercase, while others were in lowercase—were identified and addressed to standardize and clean the dataset before further processing and model training.
c. *Handling Mising Values:* The SimpleImputer from sklearn.impute was used to handle missing values by replacing them with the mean of each column. We set the imputer's strategy to 'mean', then applied the fit_transform method to compute column means and fill in missing values. For example, in a column {A} with values {[1, 2, NaN, 4, 5]}, the missing value (NaN) would be replaced by the mean, resulting in {[1, 2, 3, 4, 5]}. This ensures no data is lost due to missing values.
d. *Feature Engineering*: To enhance pattern recognition, new features such as year-over-year

ISBN 978-984-35-5270-9

rainfall changes, monthly rainfall ratios, and seasonal indicators were introduced, expanding the dataset from 14 to 43 features.

e. *Data Splitting:* The dataset was split into 60% for training and 40% for testing, ensuring the model was both well-trained and properly validated.

f. *Feature Scaling:* Scikit-learn's StandardScaler was used to standardize the data. An instance of StandardScaler was first created, which computes the mean and standard deviation necessary for scaling each categorical feature of the dataset. The categorical features were then transformed using the StandardScaler() method, ensuring that each feature had a mean of 0 and a standard deviation of 1. This standardization process helps improve the performance and stability of machine learning models.

g. *Data Augmentation:* Data augmentation enhances the model's ability to generalize by exposing it to more varied inputs and conditions. In the context of rainfall prediction and flood forecasting, these techniques can create diverse scenarios for training, improving the model's robustness to real-world unpredictability. In this case, SMOTE (Synthetic Minority Over-sampling Technique) was used, a popular data augmentation method for imbalanced classification problems. While SMOTE is typically employed to address class imbalances by generating synthetic examples of the minority class, it can also be applied to time series or regression tasks in certain ways to augment the dataset.

Applying Models: Five machine learning classification models were employed to predict flood events and their performance compared:

Random Forest (RF): A robust ensemble method that builds multiple decision trees and combines their outputs. It delivered the best performance, achieving 92% accuracy and 93% F1-score, effectively handling non-linearity and noise in the data.

Decision Tree: A straightforward, interpretable model that splits data based on feature thresholds. While fast and easy to understand, it may overfit on smaller datasets.

K-Nearest Neighbors (KNN): A distance-based algorithm that classifies each instance by a majority vote among its closest neighbors. It performs well on well-structured data but is sensitive to feature scaling and noise.

Gradient Boosting (GB): An advanced ensemble technique that builds models sequentially, each correcting the errors of the previous. It offers strong predictive performance but is more computationally intensive.

Logistic Regression (LR): A baseline linear model used for binary classification. It is simple and efficient but showed limited effectiveness in capturing complex flood patterns in this study.

Tuning Hyperparameters: Hyperparameter tuning is a crucial step in optimizing machine learning models. It involves systematically searching for the best combination of hyperparameter values that maximize the model's performance. Here, we use Grid Search which is one of the most popular and straightforward techniques for hyperparameter tuning.

10) Model Evaluation: Model evaluation is a critical part of assessing how well a machine learning (ML) model performs in predicting flood events, hydrographs, or flood-prone areas. Evaluation helps determine whether a model is reliable, accurate, and suitable for deployment in real-world flood forecasting.

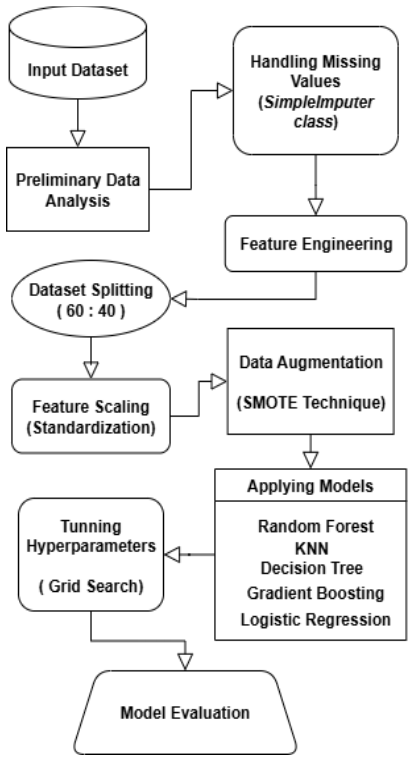

Fig. 1. Methodology of the Proposed System.

III RESULTS

1. Flood Criteria Analysis: Rainfall in Chattogram peaks between June and August, aligning with the monsoon season. Analysis revealed that annual rainfall above 3000mm and June-September totals over 2000mm are strongly linked to flood events. These thresholds serve as key indicators for flood risk assessment and early warning.

2. Performance Evaluation: Among five models, Random Forest: (Accuracy~92%), F1-score 0.93; strong performance in all metrics.

ISBN 978-984-35-5270-9 https://icebtm.iub.edu.bd

Gradient Boosting: (Accuracy~83%), F1-score 0.83; consistent but slightly less effective.
KNN and Decision Tree: (Accuracy~83%); moderate performance.
Logistic Regression: (Accuracy~75%); weakest classifier.

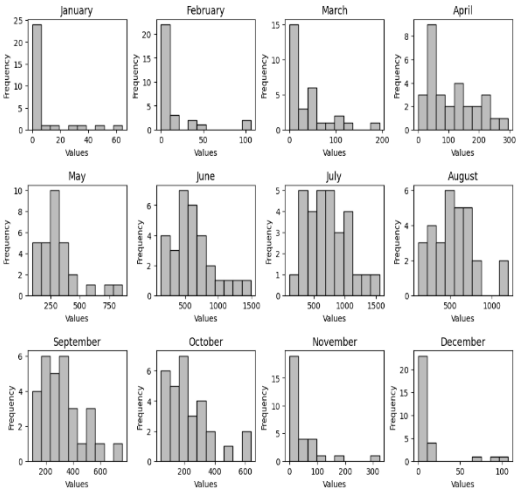

Fig. 2. Average Rainfall

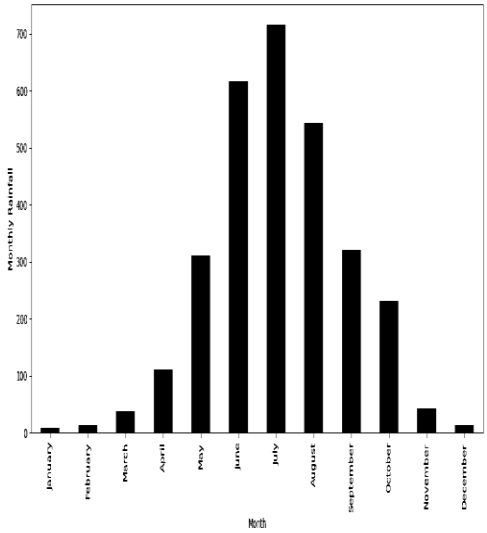

Fig. 3. Rainfall Distribution

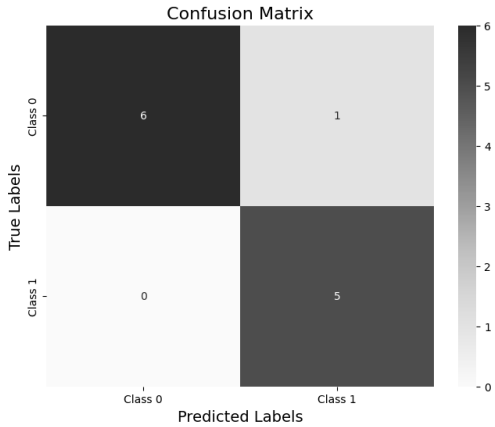

Fig. 4. Confusion Matrix of Random Forest Classifier

IV DISCUSSION

This study highlights clear rainfall trends in Chattogram, with peak monsoon rainfall (>3000 mm annually) strongly linked to flood occurrences. Using machine learning models—especially Random Forest with 92% accuracy—enabled accurate flood prediction based on historical data. Technologically, this data-driven approach enhances real-time risk forecasting and supports decision-making. From a sustainability perspective, it helps in designing early warning systems, adaptive infrastructure, and informed urban planning, directly contributing to climate resilience (SDGs 11 & 13).The method is scalable and can be applied to other flood-prone regions, offering a valuable tool for integrating environmental data and predictive modeling in disaster risk reduction.

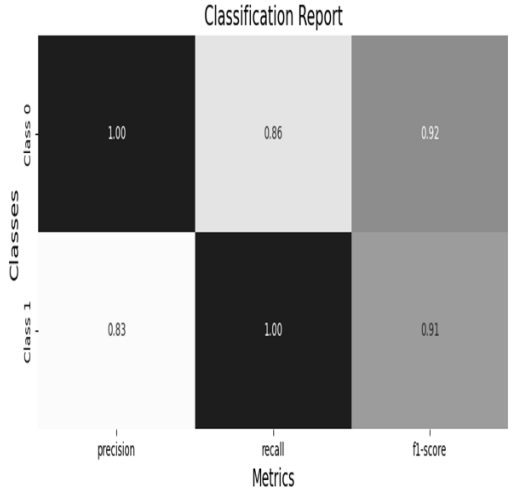

Fig. 5. Classification Report of Random Forest Classifier

ISBN 978-984-35-5270-9

V. CONCLUSION

This study presents a data-driven framework for flood prediction in the Chattogram district, leveraging historical rainfall data and advanced machine learning techniques. Among the tested models, the Random Forest classifier emerged as the most accurate, achieving a 92% accuracy and a F1-score of 0.93, effectively capturing the complex relationships between rainfall patterns and flood events. The analysis identified critical rainfall thresholds—annual rainfall exceeding 3000 mm and June–September totals above 2000 mm—as strong predictors of flood risk. These insights, coupled with robust data preprocessing and feature engineering, highlight the potential of integrating meteorological data with machine learning for improved early warning systems and climate resilience. The methodology not only enhances flood preparedness but also offers a scalable template for future applications in similar flood-prone regions.

Limitations: Small dataset size, limited regional scope, and absence of real-time integration.

Future Work: Use of real-time IoT sensors, CNN/RNN architectures, uncertainty quantification, and app-based deployment for public flood alerts.

REFERENCES

[1]. Lee, D. Perera, T. Glickman, and L. Taing, "Water-related disasters and their health impacts: A global review," *Progress in Disaster Science*, vol. 8, p. 100123, 2020.

[2]. Bangladesh Meteorological Department, “Bangladesh meteorological department live data portal,” 2024. [Online]. Available:https://live6.bmd.gov.bd/

[3]. Z. Zhu and Y. Zhang, “Flood disaster risk assessment based on random forest algorithm,” Neural Computing and Applications, vol. 34, no. 5, pp. 3443–3455, 2022

[4]. A. Y. Felix and T. Sasipraba, “Flood detection using gradient boost machine learning approach,” in Proc. Int. Conf. Computational Intelligence and Knowledge Economy (ICCIKE), 2019, pp. 779–783.

[5]. S. V. Razavi-Termeh et al., “Enhancing flood-prone area mapping: Fine-tuning the k-nearest neighbors (KNN) algorithm for spatial modelling,” Int. J. Digital Earth, vol. 17, no. 1, 2024

[6]. S. V. Razavi-Termeh et al., “Enhancing flood-prone area mapping: Fine-tuning the k-nearest neighbors (KNN) algorithm for spatial modelling,” Int. J. Digital Earth, vol. 17, no. 1, 2024.

[7]. G. Jagannathan, K. Pillaipakkamnatt, and R. N. Wright, “A practical differentially private random decision tree classifier,” in Proc. IEEE Int. Conf. Data Mining Workshops, 2009, pp. 114–121.

[8]. A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array*, vol. 16, p. 100258, 2022, doi: 10.1016/j.array.2022.100258.

[9]. K. Nafee, M. S. Al Fahad, M. K. I. Tuhin, M. S. Hossen, and M. S. Ullah, "Mapping of landslide susceptibility using state-of-the-art method and geospatial techniques in the Rangamati district in the Chattogram hill tracts region of Bangladesh," in *Landslide: Susceptibility, risk assessment and sustainability: Application of geostatistical and geospatial modeling*, pp. 103–152, Springer, 2024.

[10]. A. Omar and T. Abd El-Hafeez, "Optimizing epileptic seizure recognition performance with feature scaling and dropout layers," *Neural Computing and Applications*, vol. 36, no. 6, pp. 2835–2852, 2024.

ISBN 978-984-35-5270-9