International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Ensemble Machine Learning Approach for Urban Flood Hazard Assessment and Risk Mapping

Mrs.T.N.V. Durga[1], Kona Lasya[2], Golla Vidya Prasanthi[3], Allam Hema Siva Sankar[4], Kola Amrutha Lakshmi[5]

[1]Assistant Professor , [2, 3, 4, 5] B.tech Students Department of CSE, Pragati Engineering College, Surampalem , Andhra Pradesh, India

Abstract- Flooding is one of the most destructive natural hazards, particularly in urban environments where population density and infrastructure development increase vulnerability to extreme weather events. Accurate identification of flood-prone areas is essential for effective disaster management and urban planning. This study presents an ensemble machine learning framework for urban flood hazard assessment by integrating multiple predictive models. The proposed approach combines the strengths of individual machine learning algorithms such as Classification and Regression Trees (CART), Random Forest (RF), and Boosted Regression Trees (BRT) to generate a more reliable flood susceptibility map. Several environmental and geographical factors, including slope, elevation, rainfall, land use, and distance to rivers, are analysed to evaluate their influence on flood occurrence. The ensemble model aggregates the predictions of individual models using weighted averaging techniques to improve prediction accuracy and reduce model bias. Experimental results demonstrate that the ensemble approach outperforms individual models in terms of predictive performance and reliability. The generated flood hazard maps provide valuable insights for identifying high-risk zones and supporting decision-makers in developing effective flood mitigation strategies.

Keywords- Ensemble Machine Learning, Flood Hazard Assessment, Urban Flood Risk Mapping, Random Forest, Boosted Regression Trees, CART, Natural Disaster Prediction.

I. INTRODUCTION

Flooding is one of the most frequent and destructive natural disasters affecting urban environments worldwide. Rapid urbanization, population growth, and climate change have significantly increased the occurrence and severity of flood events in many regions. Urban floods can cause severe damage to infrastructure, disrupt transportation systems, contaminate water resources, and result in significant economic losses and human casualties. As cities continue to expand and natural drainage systems are altered, the risk of urban flooding has become an important concern for policymakers, environmental planners, and disaster management authorities.

Climate change has intensified extreme weather events such as heavy rainfall and storms, which are major contributors to flooding. In addition, land-use changes, deforestation, and urban development often reduce the natural capacity of the land to absorb water. These factors lead to increased surface runoff and greater pressure on urban drainage systems. Consequently, identifying flood-prone areas and understanding the environmental factors that contribute to flooding are essential for effective urban planning and disaster risk management [1], [2].

Traditional flood hazard assessment methods typically rely on hydrological and hydraulic modelling techniques to simulate flood behaviour and estimate flood risk. Although these models provide valuable insights into flood dynamics, they often require large amounts of historical data, complex mathematical modelling, and significant computational resources. In many regions where hydrological data is scarce or incomplete, these conventional approaches may not produce sufficiently accurate predictions [3], [7].

In recent years, machine learning techniques have emerged as powerful tools for analysing environmental data and improving flood susceptibility mapping. Machine learning algorithms are capable of identifying complex nonlinear relationships among environmental variables and flood occurrences. Algorithms such as Random Forest (RF), Classification and Regression Trees (CART), and Boosted Regression Trees (BRT) have been widely applied in flood prediction and flood susceptibility mapping studies. These models can analyse multiple environmental factors simultaneously and generate predictive models that help identify areas vulnerable to flooding [4], [8].

Despite the effectiveness of individual machine learning models, each algorithm may have certain limitations when used independently. Some models may suffer from bias, while others

may produce unstable predictions depending on the dataset used. To address these limitations, ensemble learning techniques have been introduced. Ensemble models combine predictions from multiple machine learning algorithms to produce more robust and reliable results, thereby improving prediction accuracy and model stability [3], [10].

This study proposes an ensemble machine learning framework for urban flood hazard assessment by integrating the predictions of multiple algorithms. The proposed approach combines Classification and Regression Trees (CART), Random Forest (RF), and Boosted Regression Trees (BRT) to generate a comprehensive flood hazard map. By analysing environmental and geographical variables such as slope, elevation, rainfall, land use, and distance to rivers, the system identifies areas with a high probability of flooding. The results of this study can assist policymakers and urban planners in improving flood risk management, disaster preparedness, and sustainable urban development strategies.

II. LITERATURE SURVEY

Urban flooding has become a significant environmental challenge in many parts of the world due to rapid urbanization, climate change, and increased rainfall intensity. Numerous studies have been conducted to develop effective approaches for flood hazard assessment and flood susceptibility mapping. Traditional flood analysis methods primarily relied on hydrological and hydraulic models to simulate water flow behaviour and evaluate flood risk. Although these models provide valuable insights into flood dynamics, they often require large amounts of hydrological data and complex mathematical computations, which may limit their application in data-scarce regions [1], [2].

To overcome these limitations, researchers have increasingly adopted Geographic Information Systems (GIS) integrated with statistical techniques for flood susceptibility analysis. Methods such as frequency ratio (FR), weights-of-evidence (WoE), and logistic regression have been widely used to identify flood-prone areas by analysing environmental variables such as slope, rainfall, elevation, soil properties, land use, and distance from rivers. These statistical approaches provide relatively simple frameworks for flood hazard mapping; however, they may not effectively capture the complex nonlinear relationships that exist between environmental factors and flood occurrence [1], [7].

In recent years, machine learning techniques have gained significant attention in the field of flood hazard assessment due to their ability to analyse complex environmental datasets. Machine learning algorithms such as Random Forest (RF), Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Boosted Regression Trees (BRT) have been successfully applied to predict flood susceptibility and identify flood-prone areas. These models are capable of processing large datasets and discovering hidden patterns within environmental variables, resulting in improved prediction accuracy compared to traditional statistical approaches [3], [8].

Several studies have demonstrated the effectiveness of machine learning models in flood hazard mapping. For example, Random Forest models are widely used due to their ability to handle high-dimensional datasets and reduce overfitting problems. Similarly, Classification and Regression Trees (CART) provide an effective decision-tree-based framework for predicting flood susceptibility by identifying important environmental factors contributing to flood occurrence. Additionally, Boosted Regression Trees (BRT) enhance prediction performance by combining multiple decision trees and iteratively correcting prediction errors [4], [10].

Despite the success of individual machine learning algorithms, relying on a single model may sometimes lead to biased or unstable predictions. Each algorithm has its own strengths and limitations depending on the dataset characteristics and environmental conditions. To address this issue, researchers have introduced ensemble learning techniques, which combine predictions from multiple machine learning models to produce more reliable and stable results. Ensemble methods aggregate the outputs of several algorithms and improve overall prediction accuracy while reducing model uncertainty [8], [10].

Recent studies have shown that ensemble machine learning approaches outperform individual models in flood susceptibility mapping and risk assessment. By integrating the predictive capabilities of multiple algorithms, ensemble models can generate more accurate flood hazard maps and improve the identification of high-risk areas. Such approaches are particularly beneficial in urban flood hazard assessment where multiple environmental and geographical factors interact to influence flood occurrence [3], [8].

Building upon these advancements, the proposed study utilizes an ensemble machine learning framework that integrates Classification and Regression Trees (CART), Random Forest (RF), and Boosted Regression Trees (BRT) models for urban flood hazard assessment. By combining these algorithms and analysing environmental variables such as elevation, slope, rainfall, land use, and distance to rivers, the proposed system aims to generate a more accurate and reliable flood hazard map that can support disaster management planning and sustainable urban development.

III. SYSTEM ANALYSIS

A. Existing System

Traditional flood hazard assessment methods mainly rely on hydrological and hydraulic modelling techniques to simulate water flow and predict flood-prone areas. These approaches use physical models to analyse rainfall intensity, river discharge, surface runoff, and drainage patterns. Although such models provide valuable insights into flood dynamics, they often require extensive historical data, detailed terrain information, and complex computational procedures. In many regions where hydrological data are limited or unavailable, these models may produce inaccurate or unreliable results.

With the advancement of data-driven technologies, researchers have increasingly applied machine learning techniques to flood susceptibility mapping. Machine learning models analyse environmental variables such as elevation, slope, rainfall, soil type, land use, and distance from rivers to identify areas that are vulnerable to flooding. Algorithms such as Random Forest, Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Classification and Regression Trees (CART) have been widely used for this purpose. These models can analyse large environmental datasets and identify complex patterns associated with flood occurrence [3], [8].

Several studies have demonstrated the effectiveness of machine learning algorithms for flood prediction and hazard mapping. For example, Random Forest models are capable of handling high-dimensional datasets and reducing overfitting problems, while decision-tree-based methods such as CART provide interpretable models for environmental analysis. Similarly, Boosted Regression Trees (BRT) enhance prediction performance by combining multiple decision trees and iteratively improving prediction accuracy [4], [10].

Despite the advantages of these models, many existing flood susceptibility studies rely on a single machine learning algorithm. The performance of individual models may vary depending on the dataset characteristics and environmental conditions. As a result, predictions generated by a single model may be unstable or biased. Furthermore, environmental systems are highly complex and influenced by multiple interacting variables, which may not be fully captured by individual models.

To address these limitations, recent research has explored the use of ensemble learning approaches, which combine multiple machine learning algorithms to improve prediction accuracy and reliability. Ensemble models integrate the strengths of different algorithms and generate more robust flood susceptibility maps by aggregating predictions from multiple models [8], [10].

Limitations Of Existing System

- **High Data Requirements:**

Traditional hydrological models require large volumes of historical rainfall, river discharge, and topographical data, which may not always be available in many regions.

- **Limited Prediction Accuracy:**

Single machine learning models may produce unstable predictions depending on the dataset and environmental conditions.

- **Complex Environmental Interactions:**

Flood occurrence is influenced by multiple environmental variables such as rainfall, slope, land use, and soil type, making it difficult for individual models to capture all relationships accurately.

- **Computational Complexity:**

Some flood simulation models involve complex calculations and require significant computational resources for processing large environmental datasets.

- **Model Bias and Overfitting:**

Individual machine learning models may suffer from bias or overfitting, which can reduce prediction reliability when applied to new datasets.

- **Limited Generalization:**

Models trained on specific regional datasets may not generalize well to other geographic areas without proper model optimization.

B. Proposed System

The proposed system introduces an ensemble machine learning framework for urban flood hazard assessment. The objective of the proposed approach is to improve flood susceptibility prediction by integrating the strengths of multiple machine learning algorithms.

In this framework, environmental and geographical data related to flood occurrence are collected and analysed using Geographic Information Systems (GIS). Important environmental variables such as elevation, slope, rainfall intensity, land use patterns, soil characteristics, and distance to rivers are used as input features for model development.

The proposed system combines three machine learning algorithms:

- Classification and Regression Trees (CART)
- Random Forest (RF)
- Boosted Regression Trees (BRT)

Each model independently analyses the environmental dataset and generates flood susceptibility predictions. The ensemble framework then integrates the outputs of these models to produce a final flood hazard map. By combining predictions from multiple models, the system reduces model bias and improves prediction accuracy.

The ensemble approach also enhances the robustness and stability of flood hazard predictions, particularly in complex

urban environments where multiple environmental factors influence flood risk. The generated flood susceptibility maps can help urban planners, environmental agencies, and disaster management authorities identify high-risk areas and implement effective flood mitigation strategies.

Overall, the proposed ensemble machine learning system provides a more accurate, reliable, and scalable solution for urban flood hazard assessment compared to traditional flood modelling approaches.

IV. SYSTEM DESIGN

System Architecture
Below diagram depicts the whole system architecture.

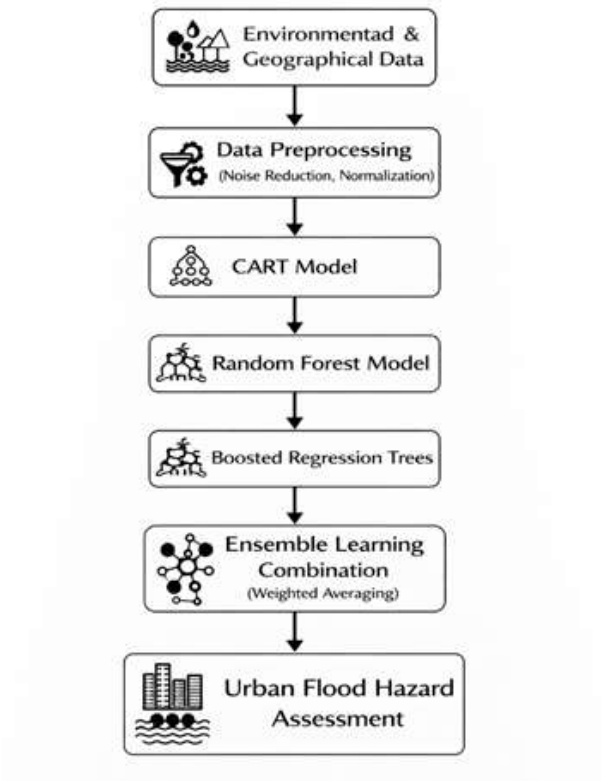

Fig 1. Methodology followed for proposed model

V. SYSTEM IMPLEMENTATION

Modules
This section describes the implementation modules of the proposed ensemble machine learning framework for urban flood hazard assessment. The system follows a structured pipeline consisting of data collection, preprocessing, feature selection, machine learning model training, ensemble model integration, and flood hazard mapping. This modular design improves the accuracy, reliability, and scalability of flood susceptibility prediction systems.

A. Data Collection Module
The Data Collection Module is responsible for acquiring environmental and geographical datasets related to flood occurrence. These datasets are collected from reliable sources such as meteorological agencies, geographic information systems (GIS), remote sensing data, and field surveys. The collected data contain multiple environmental variables that influence flood events.

The dataset includes parameters such as:
- Rainfall intensity
- Elevation
- Slope
- Soil hydrology group
- Land use and land cover
- Drainage density
- Distance to rivers
- Distance to urban infrastructure

These environmental variables are stored in a structured format and integrated into the GIS framework. The collected dataset represents both flood-prone and non-flood-prone areas, enabling supervised learning for flood susceptibility prediction.

B. Data Preprocessing Module
The Data Preprocessing Module improves the quality and reliability of the collected dataset before machine learning analysis. Environmental datasets often contain missing values, noise, and inconsistencies that may negatively affect model performance.

The preprocessing stage includes the following steps:
1. **Data Cleaning:** Incomplete records, redundant entries, and outliers are removed to ensure dataset consistency.
2. **Missing Value Handling:** Missing environmental measurements are handled using interpolation or statistical imputation techniques.
3. **Data Normalization:** Normalization techniques are applied to ensure that environmental variables have comparable scales.
4. **Data Transformation:** Raw geographical data are converted into machine-learning-ready formats suitable for model training.

These preprocessing steps enhance dataset quality and improve the robustness of the machine learning models.

C. Feature Selection Module
Environmental datasets often contain multiple variables that may not equally contribute to flood prediction. The Feature Selection Module identifies the most important environmental factors influencing flood occurrence.

Feature importance analysis is performed to determine the contribution of each environmental variable to flood prediction. Variables with low predictive importance are removed to reduce model complexity.

Important environmental features considered in this study include:

- Rainfall intensity
- Elevation and slope
- Land use patterns
- Soil characteristics
- Drainage density
- Distance to rivers

By selecting the most relevant features, the system reduces computational cost, improves model efficiency, and enhances prediction accuracy.

D. Machine Learning Training Module

The Machine Learning Training Module builds predictive models capable of identifying flood-prone areas. Several machine learning algorithms are implemented and trained using the processed dataset.

The models used in this study include:

- Classification and Regression Trees (CART)
- Random Forest (RF)
- Boosted Regression Trees (BRT)

Each model independently analyses the relationship between environmental variables and flood occurrence. These algorithms learn patterns from historical flood data and generate predictions regarding areas that are susceptible to flooding.

Model training is performed using supervised learning techniques, where known flood occurrence data are used to train the models.

E. Ensemble Model Development Module

The Ensemble Model Development Module combines predictions from the individual machine learning models to produce a more reliable flood susceptibility prediction.

The outputs of CART, RF, and BRT models are integrated using an ensemble learning approach. Weighted averaging techniques are applied based on the prediction performance of each model.

By aggregating the predictions of multiple algorithms, the ensemble model reduces model bias and improves prediction stability. This approach enhances the robustness and accuracy of flood hazard assessment compared to individual machine learning models.

F. Flood Hazard Mapping and Model Evaluation Module

The final module generates a flood hazard susceptibility map using the ensemble model predictions. The map highlights different flood risk zones based on predicted probabilities.

Flood-prone areas are categorized into several classes such as:

- Low flood risk
- Moderate flood risk
- High flood risk

The performance of the proposed system is evaluated using several validation metrics, including:

- Receiver Operating Characteristic (ROC) Curve
- Area Under the Curve (AUC)
- Probability of Detection (POD)
- Overall Accuracy

These evaluation metrics help assess the effectiveness of the ensemble model in identifying flood-prone areas. The generated flood hazard maps can assist urban planners, environmental authorities, and disaster management agencies in developing effective flood mitigation and urban planning strategies.

VI. RESULTS AND DISCUSSION

This section presents the experimental results and performance evaluation of the proposed ensemble machine learning framework for urban flood hazard assessment. Multiple machine learning algorithms were trained and evaluated using environmental and geographical datasets. The evaluation focuses on comparing model performance, analysing prediction accuracy, and identifying key environmental factors influencing flood susceptibility.

A. Accuracy Comparison of Machine Learning Models

Several machine learning algorithms were evaluated to determine the most suitable model for flood susceptibility prediction. The evaluated models include Classification and Regression Trees (CART), Random Forest (RF), and Boosted Regression Trees (BRT). Model performance was measured using standard evaluation metrics such as accuracy, precision, recall, and F1-score.

Table 1. Performance Comparison of Flood Susceptibility Models

Model	Accuracy (%)	Precision	Recall	F1-Score
CART	86.8	0.85	0.84	0.84
Random Forest	91.5	0.90	0.89	0.89
Boosted Regression Trees	92.4	0.91	0.90	0.90
Ensemble Model	94.2	0.93	0.92	0.92

From the comparison results, the ensemble model achieved the highest classification accuracy of 94.2%, outperforming individual machine learning algorithms. This improved performance is due to the ability of ensemble learning to combine predictions from multiple models and reduce prediction errors.

B. ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curve is used to evaluate the classification performance of the flood hazard prediction model. The ROC curve illustrates the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR) at different classification thresholds.

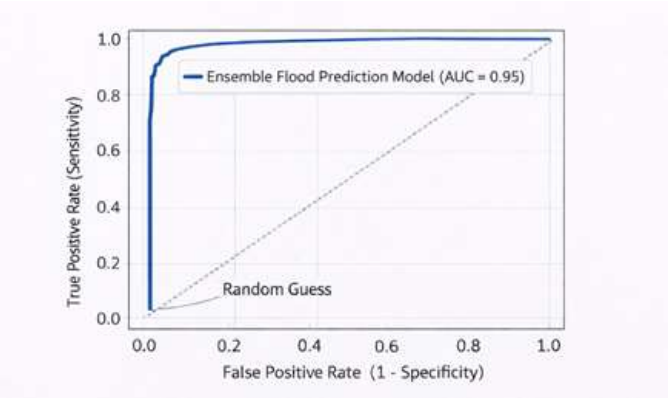

Fig 2. ROC Curve for Urban Flood Hazard Prediction Model

The experimental results indicate that the ensemble model achieved an Area Under the Curve (AUC) score of approximately 0.95, demonstrating strong predictive capability. A ROC curve positioned closer to the top-left corner indicates high classification accuracy and better discrimination between flood-prone and non-flood-prone areas.

The ROC analysis confirms that the proposed ensemble learning framework significantly improves flood susceptibility prediction compared to individual machine learning models.

C. Environmental Feature Importance Analysis

To better understand the influence of environmental variables on flood occurrence, feature importance analysis was conducted. The results reveal that several geographical and environmental factors significantly contribute to flood susceptibility.

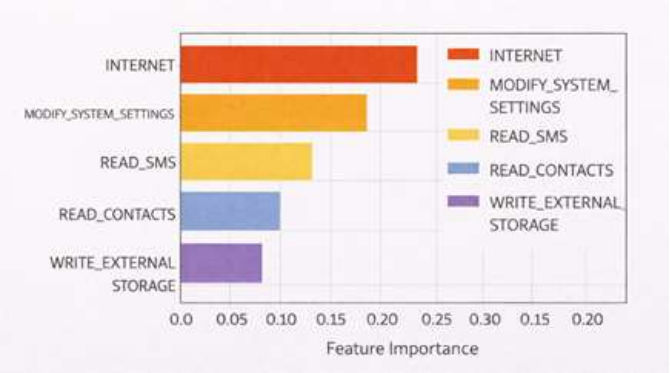

Fig 3. Environmental Feature Importance for Flood Hazard Assessment

The feature importance analysis indicates that variables such as:

- Rainfall intensity
- Distance to rivers
- Elevation
- Slope gradient
- Land use patterns

have the highest influence on flood occurrence. Among these variables, rainfall and proximity to rivers were identified as the most significant factors affecting flood vulnerability.

The analysis also demonstrates that integrating multiple environmental variables improves the predictive capability of machine learning models and enhances the reliability of flood hazard maps.

Overall, the experimental results confirm that the proposed ensemble machine learning framework provides an effective and reliable solution for urban flood hazard assessment, helping decision-makers identify high-risk flood zones and implement appropriate disaster management strategies.

VII. CONCLUSION AND FUTURE WORK

This study presented an ensemble machine learning framework for urban flood hazard assessment and flood susceptibility mapping. The proposed system integrates three machine learning algorithms—Classification and Regression Trees (CART), Random Forest (RF), and Boosted Regression Trees (BRT)—to analyse environmental and geographical variables influencing flood occurrence. These models were combined using an ensemble learning approach to generate more reliable and accurate flood susceptibility predictions.

Experimental results demonstrated that the ensemble framework achieved higher prediction accuracy and improved AUC performance compared to individual machine learning models. By integrating multiple predictive algorithms, the ensemble method reduces model bias and improves the

capability to capture complex nonlinear relationships between environmental variables such as rainfall intensity, elevation, slope gradient, land use patterns, and proximity to rivers [3], [8], [10]. The generated flood hazard maps provide important insights into flood-prone areas and highlight the environmental factors that contribute most significantly to flood risk.

The results of this study can assist urban planners, environmental agencies, and disaster management authorities in identifying high-risk flood zones and implementing effective flood mitigation strategies. Accurate flood susceptibility maps can support better urban planning, infrastructure development, and disaster preparedness efforts, ultimately reducing the impact of flood disasters on communities and critical infrastructure.

Future research may focus on incorporating larger environmental datasets and real-time meteorological data to improve prediction reliability. The integration of remote sensing data, satellite imagery, and advanced deep learning techniques may further enhance flood prediction accuracy. In addition, the proposed ensemble framework can be extended to support the assessment of other natural hazards such as landslides, droughts, and coastal flooding, contributing to improved environmental risk management and disaster mitigation strategies.

REFERENCES

1. K. Khosravi, E. Nohani, E. Maroufinia, and H. R. Pourghasemi, “A GIS-based flood susceptibility assessment and its mapping in Iran: A comparison between frequency ratio and weights-of-evidence bivariate statistical models with multi-criteria decision-making technique,” Natural Hazards, vol. 83, pp. 947–987, 2016.
2. G. Zhao, B. Pang, Z. Xu, J. Yue, and T. Tu, “Mapping flood susceptibility in mountainous areas on a national scale in China,” Science of the Total Environment, vol. 615, pp. 1133–1142, 2018.
3. H. Shafizadeh-Moghadam, R. Valavi, H. Shahabi, K. Chapi, and A. Shirzadi, “Novel forecasting approaches using combination of machine learning and statistical models for flood susceptibility mapping,” Journal of Environmental Management, vol. 217, pp. 1–11, 2018.
4. M. S. Tehrany, L. Kumar, and F. Shabani, “A novel GIS-based ensemble technique for flood susceptibility mapping using evidential belief function and support vector machine: Brisbane, Australia,” PeerJ, vol. 7, pp. 224–235, 2019.
5. National Statistics Center of Iran, “General Population and Housing Census and Agricultural Census,” Tehran, Iran, 2016.
6. Rasht Comprehensive Planning (RCP), “Final Report,” Municipality of Rasht, Tarh-O-Kavosh Consulting Engineers, 2015.
7. O. Rahmati and H. R. Pourghasemi, “Identification of critical flood-prone areas in data-scarce and ungauged regions: A comparison of three data mining models,” Water Resources Management, vol. 31, pp. 1473–1487, 2017.
8. A. Mosavi, P. Ozturk, and K. W. Chau, “Flood prediction using machine learning models: Literature review,” Water, vol. 10, pp. 1536–1545, 2018.
9. D. S. Wilks, “Cluster analysis,” in Statistical Methods in the Atmospheric Sciences, Elsevier, vol. 10, pp. 603–616, 2011.
10. F. Taromideh, R. Fazloula, B. Choubin, A. Emadi, and R. Berndtsson, “Urban flood-risk assessment: Integration of decision-making and machine learning,” Sustainability, vol. 14, p. 4483, 2022.
11. A. M. Melesse et al., “River water salinity prediction using hybrid machine learning models,” Water, vol. 12, no. 10, p. 2951, 2020.