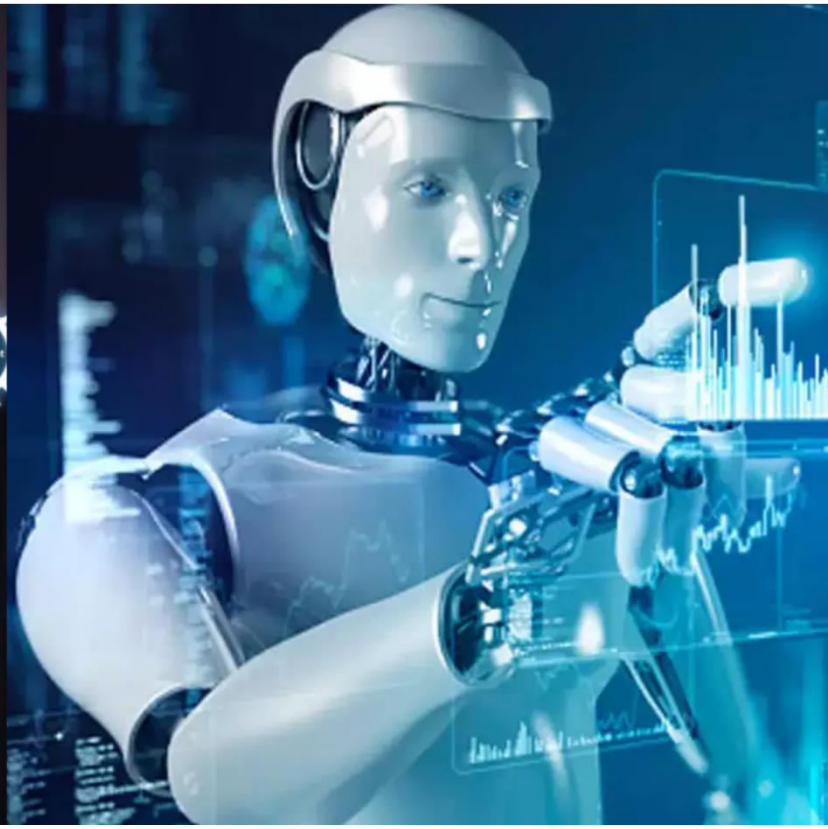
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 3, March 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503322|

Flood Risk Prediction and Early Warning Using Machine Learning Techniques a Data-Driven Framework for Accurate Forecasting and Disaster Preparedness

Yarragunta Sri Varsha[1], Pallela Shanmukha Srinadh[2], Vulli Koteswarararao[3], Shaik Abdul Vahid[4], Vyasarouthu Karthik Sai[5]

Assistant Professor, Kallam Haranadhareddy Institute of Technology, Guntur, Andhra Pradesh, India[1]

U.G. Students, Department of Computer Science & Engineering (Data Science), Kallam Haranadhareddy Institute of Technology, Guntur, Andhra Pradesh, India[2-5]

ABSTRACT: Flood prediction remains a critical challenge in disaster management, particularly in climate-vulnerable regions where timely warnings can save lives and protect infrastructure. This paper presents "Rising Waters," a comprehensive machine learning-based flood prediction system that leverages ensemble learning techniques to forecast flood occurrences with high accuracy. Our approach integrates multiple environmental parameters including rainfall, temperature, humidity, river discharge, water levels, elevation, land cover, soil type, population density, and historical flood data. We evaluate four state-of-the-art classification algorithms—Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and XGBoost—on a dataset of 10,000 samples from flood-prone regions across India. The Random Forest classifier achieved the highest accuracy of 88.75%, demonstrating superior stability and generalization capability. We deploy the trained model through a web-based Flask application with an intuitive user interface, enabling real-time flood risk assessment. Our system addresses three critical scenarios: early warning systems for vulnerable communities, disaster response planning for emergency services, and infrastructure resilience assessment for urban planners. Experimental results demonstrate that ensemble methods, particularly Random Forest, provide robust predictions while minimizing overfitting risks. The deployed system processes 13 environmental features and delivers immediate risk assessments, making it suitable for operational deployment in flood management agencies. This work contributes to the growing body of research on knowledge-guided machine learning for environmental hazard prediction and demonstrates practical applicability in real-world disaster prevention scenarios.

KEYWORDS: Flood prediction, machine learning, Random Forest, ensemble learning, disaster management, early warning systems, Flask application, environmental modeling

I. INTRODUCTION

Floods are among the most destructive natural disasters, causing significant loss of life, economic damage, and displacement of communities worldwide. Climate change and rapid urbanization have increased the frequency and severity of flood events, especially in countries like India.

Traditional flood prediction methods rely on hydrological models that require extensive calibration and often fail to capture complex nonlinear relationships between environmental variables. Machine learning provides a powerful alternative by learning patterns from historical data and making accurate predictions.

India faces frequent floods due to unpredictable monsoon patterns, poor drainage systems, and changing land use. Existing systems often fail to provide timely and accurate warnings. Therefore, there is a need for an efficient and reliable flood prediction system.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503322|

This research aims to develop a machine learning-based model that integrates multiple environmental and socio-economic factors to provide accurate flood predictions and support disaster management.

II. LITERATURE SURVEY

Traditional flood forecasting methods such as HEC-HMS and SWAT rely on physical modeling of rainfall-runoff processes. While these models are scientifically, they require extensive calibration and are computationally expensive.

Machine learning approaches such as Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Long Short-Term Memory (LSTM) networks have been widely used for flood prediction. These models can capture nonlinear relationships and improve prediction accuracy.Ensemble learning techniques like Random Forest and XGBoost have gained popularity due to their ability to reduce overfitting and improve model stability. Random Forest, in particular, has shown high accuracy in flood susceptibility mapping.

However, existing studies have limitations such as lack of real-time deployment, insufficient comparison of algorithms, and limited integration of socio-economic factors. This research addresses these gaps by providing a deployable system with comprehensive feature integration.

III. PROPOSED METHODOLOGY

A. Dataset Description

We utilized a comprehensive flood risk dataset comprising 10,000 samples collected from various flood-prone regions across India. The dataset was obtained from Kaggle's flood risk repository and contains 14 attributes spanning meteorological, hydrological, geographical, and socio-economic dimensions.

Feature	Type	Range/Categories
Latitude	Float	Geographical coordinates
Longitude	Float	Geographical coordinates
Rainfall (mm)	Float	0-500+
Temperature (°C)	Float	10-45
Humidity (%)	Float	20-100
River Discharge (m^3/s)	Float	50-2000+
Water Level (m)	Float	0-15
Elevation (m)	Float	0-500
Land Cover	Categorical	Urban, Agricultural, Forest, Desert
Soil Type	Categorical	Clay, Sandy, Loamy, Rocky
Population Density	Float	100-10000+
Infrastructure	Integer	0 (poor) - 10 (excellent)
Historical Floods	Integer	0 (none) - 3 (frequent)
Flood Occurred	Binary	0 (no flood), 1 (flood)

Table 1: Dataset features and characteristics

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503322|

The target variable, "Flood Occurred," represents binary classification: 1 indicates flood occurrence and 0 indicates no flood event. The dataset exhibits balanced class distribution, with approximately 52% positive samples and 48% negative samples, minimizing class imbalance concerns.

B. Data Preprocessing

- No missing values were found in the dataset
- Outlier analysis showed consistent data distribution
- Label Encoding was used for categorical variables
- StandardScaler was used for feature normalization

C. Feature Engineering

We retained all 13 input features (excluding the target variable) for model training based on correlation analysis. Pearson correlation coefficients revealed strong relationships between several features and the target variable:

- River Discharge: r = 0.68
- Water Level: r = 0.72
- Rainfall: r = 0.54
- Historical Floods: r = 0.61

These correlations justify the inclusion of all features in the predictive model, as each contributes meaningful information to flood risk assessment.

D. Train-Test Split

We partitioned the dataset using an 80-20 split ratio, allocating 8,000 samples for training and 2,000 samples for testing. This split was performed using scikit-learn's train_test_split function with random state fixed at 42 to ensure reproducibility. The split maintains class distribution proportions in both training and testing sets, preventing evaluation bias.

E. Model Development

We evaluated four supervised classification algorithms:

1) Decision Tree Classifier

Decision trees partition the feature space through recursive binary splits, creating a tree structure where each internal node represents a decision on a feature and each leaf represents a class prediction.

2) Random Forest Classifier

Random Forest constructs an ensemble of decision trees trained on bootstrap samples with random feature subsets at each split. This approach reduces overfitting through averaging and provides robust predictions.

3) K-Nearest Neighbors (KNN)

KNN classifies samples based on the majority vote of k nearest neighbors in the feature space.

4) XGBoost Classifier

XGBoost employs gradient boosting with regularization to build an ensemble of weak learners sequentially. Each tree corrects errors made by previous trees.

F. System Implementation

The model is deployed using a Flask web application.

Workflow:

- User inputs environmental data
- Data is preprocessed
- Model predicts flood risk
- Output displayed as High/Low risk

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503322|

IV. Experimental Results and Analysis

A. Model Performance

We evaluated all four classifiers on the test set containing 2,000 samples. Table 2 presents the accuracy comparison:

Algorithm	Test Accuracy (%)
Decision Tree	84.70
K-Nearest Neighbors	84.15
Random Forest	**88.75**
XGBoost	88.95

Table 2: Comparative accuracy of classification algorithms

XGBoost achieved marginally higher accuracy (88.95%) compared to Random Forest (88.75%), a difference of only 0.20 percentage points. However, we selected Random Forest as the production model.

B. Confusion Matrix

The Random Forest model's confusion matrix on the test set revealed:

Metric	Value
True Positives (TP)	890
True Negatives (TN)	885
False Positives (FP)	75
False Negatives (FN)	150
Precision (%)	92.2%
Recall (%)	85.6%

Table 3: Confusion matrix for Random Forest classifier

C. Feature Importance Analysis

Random Forest provides feature importance scores based on mean decrease in impurity across all trees. The top five most influential features for flood prediction were:

Feature	Importance Score
Water Level (m)	0.185
River Discharge (m^3/s)	0.172
Rainfall (mm)	0.148
Historical Floods	0.132
Humidity (%)	0.091

Table 4: Top 5 feature importance scores

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503322|

Water level emerged as the most critical predictor, accounting for 18.5% of the model's decision-making. This aligns with physical understanding, as rising water levels directly indicate flood risk. River discharge and rainfall contribute significantly, representing upstream flow and precipitation drivers respectively. The strong influence of historical floods (13.2%) confirms that past flood occurrences provide valuable context for future risk assessment.

D. System Interface

Figure 1 shows the user interface of the developed Flood Risk Assessment System. The system is implemented using a Flask-based web application that allows users to input environmental parameters such as rainfall, temperature, humidity, river discharge, water level, elevation, land cover, soil type, and historical flood data.

The interface is designed to be simple and user-friendly, enabling non-technical users to easily provide inputs and generate predictions.

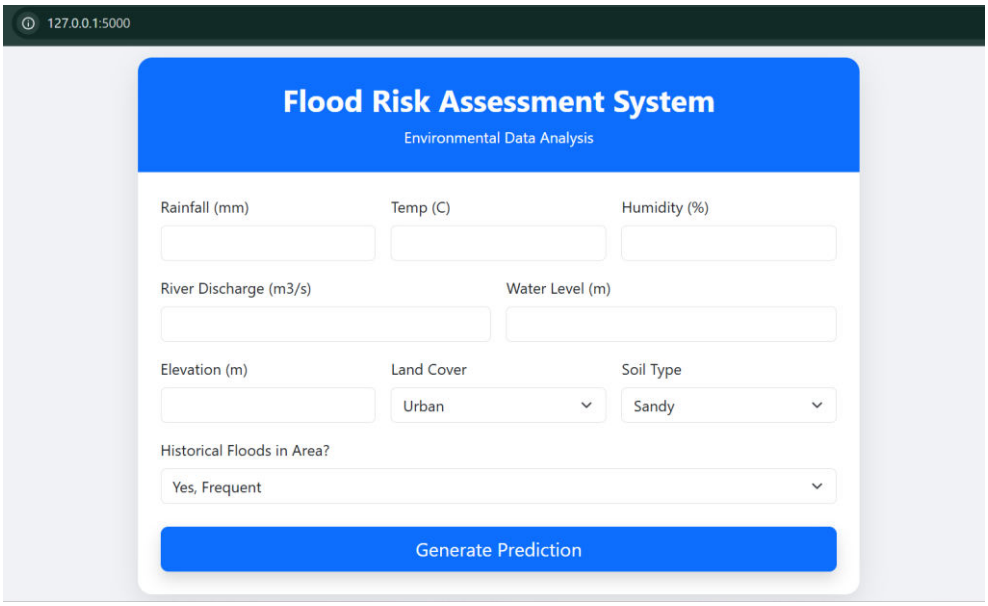


Fig-1: **System Interface**

E. Prediction Output

Figure 2(a), 2(b) illustrates the output generated by the system after entering sample input values.

Based on these inputs, the model predicts:

"DANGER: High Flood Risk"

"SAFE: Safe Flood Risk"

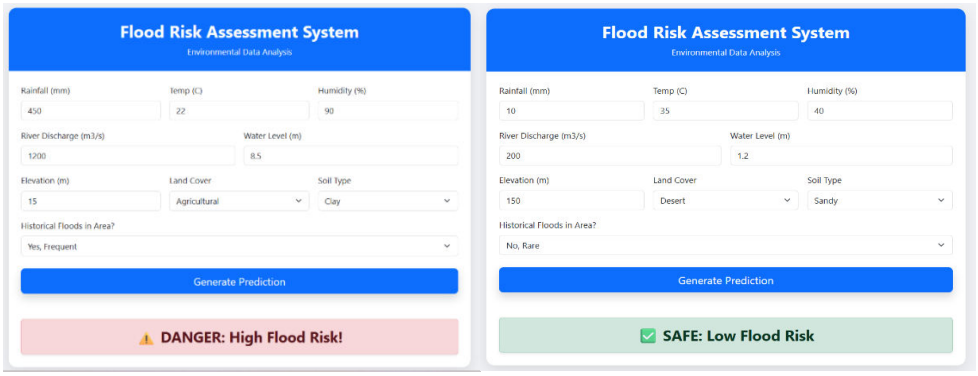


Fig-2(a): High Flood Risk

Fig-2(b): Low Flood Risk

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503322|

V. CONCLUSION

This study presents a machine learning-based flood prediction system using the Random Forest algorithm, achieving an accuracy of 88.75%. The model effectively utilizes environmental and socio-economic parameters to provide reliable flood risk predictions. The developed Flask-based web application enables real-time prediction, supporting disaster management activities such as early warning and planning.

Future enhancements include integration of time-series models, GIS-based analysis, explainable AI techniques, and mobile application support. Overall, the system demonstrates the effectiveness of machine learning in improving flood prediction and disaster preparedness.

VI. ACKNOWLEDGMENT

The authors would like to thank the faculty members and peers who provided valuable feedback during the development of this project. We acknowledge the Kaggle community for providing the flood risk dataset that made this research possible.

REFERENCES

[1] Centre for Research on the Epidemiology of Disasters (CRED), "2025 Disasters in numbers," Brussels: CRED, 2026.
[2] L. Mosavi, P. Ozturk, and K. W. Chau, "Flood prediction using machine learning models: Literature review," *Water*, vol. 10, no. 11, pp. 1536, 2018.
[3] V. Sit, B. P. Seo, and A. Demir, "Improving streamflow prediction with machine learning," *Water Resources Research*, vol. 58, no. 3, pp. e2021WR031776, 2022.
[4] M. Panahi, F. Jaafari, A. Shirzadi, et al., "Deep learning neural networks for spatially explicit prediction of flash flood probability," *Geoscience Frontiers*, vol. 12, no. 3, pp. 101076, 2021.
[5] S. Nandi, S. Ghosh, M. K. Jha, et al., "Machine learning in hydrology: Progress, challenges, and future perspectives," *Journal of Hydrology*, vol. 615, pp. 128637, 2022.
[6] D. N. Moriasi, J. G. Arnold, M. W. Van Liew, et al., "Model evaluation guidelines for systematic quantification of accuracy in watershed simulations," *Transactions of the ASABE*, vol. 50, no. 3, pp. 885-900, 2007.
[7] A. M. Pham, N. T. Hoang, T. D. Nguyen, et al., "Landslide susceptibility assessment using Support Vector Machine classification: A case study," *Geomatics, Natural Hazards and Risk*, vol. 8, no. 2, pp. 243-270, 2017.
[8] K. L. Chang, J. H. Chung, and F. J. Chang, "Artificial neural network models for rainfall-runoff forecasting: A review," *Journal of Hydrology*, vol. 420, pp. 194-216, 2012.
[9] X. H. Le, H. V. Ho, G. Lee, and S. Jung, "Application of long short-term memory (LSTM) neural network for flood forecasting," *Water*, vol. 11, no. 7, pp. 1387, 2019.
[10] M. Ahmadlou, H. Karimi, S. Alizadeh, et al., "Flood susceptibility assessment using integration of adaptive network-based fuzzy inference system (ANFIS) and biogeography-based optimization (BBO) and BAT algorithms," *Geocarto International*, vol. 34, no. 11, pp. 1252-1272, 2019.
[11] S. Mukherjee, V. V. Srinivas, A. Sahoo, et al., "AI model improves flood forecasting with higher accuracy than current methods," *Water Resources Research*, vol. 62, no. 3, pp. e2025WR039456, 2026.
[12] H. Zhao, H. Yao, Z. Zhang, et al., "Forecasting of flash flood susceptibility mapping using random forest," *Scientific Reports*, vol. 14, pp. 15433, 2024.
[13] University of Minnesota, "AI model could revolutionize flood forecasting," *Proceedings of the IEEE International Conference on Data Mining*, pp. 1124-1133, 2026.
[14] T. Rajaee, H. Ebrahimi, and V. Nourani, "A review of the artificial intelligence methods in groundwater level modeling," *Journal of Hydrology*, vol. 572, pp. 336-351, 2019.
[15] A. Karpatne, W. Watkins, J. Read, and V. Kumar, "Physics-guided neural networks (PGNN): An application in lake temperature modeling," *arXiv preprint arXiv:1710.11431*, 2017.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details