IJRCCES

International Journal for Research in Climate Change and Earth Science

Published By:
Green Publications Services Private LTD.

Volume 01 Issue 01 March 2026
DOI: 10.53555/cces.v1i1.2573

Mohan Lal 'Arya'

Professor & Dean, Department of Education, School of Education and Humanities, IFTM University, Moradabad UP, drmlarya2012@gmail.com, https://orcid.org/0000-0001-5424-8819, 244102

Corresponding Author (Mohan Lal 'Arya')

PREDICTING FLOOD RISK PROBABILITY USING ENVIRONMENTAL AND INFRASTRUCTURE-BASED INDICATORS A MACHINE LEARNING APPROACH

Article History:
Article Type: Research

Received Date: 09/02/2026 **Revised Date: 10/03/1026**
Acceptance Date: 18/03/2026 **Published Date: 25/03/2026**

Abstract

Climate change, environmental degradation, infrastructure quality, and socioeconomic vulnerability collectively shape flood risk. This study developed machine-learning models to estimate flood probability using environmental, infrastructure, and socioeconomic indicators. The data analyzed were acquired from a database with 1,117,957 labelled observations, 20 predictor variables and the continuous outcome variable, Flood Probability. Data pre-processing involved checking for missing values, duplicate values, consistency of the data and elimination of the non-informative identifier variable. Six regression algorithms Decision Tree, Random Forest, Gradient Boosting, XGBoost, LightGBM, and CatBoost were evaluated using mean absolute error, mean squared error, root mean squared error, and coefficient of determination. CatBoost was the model that had the best predictive ability among the evaluated models, with a test MAE of 0.0220, an RMSE of 0.0267, and an R^2 of 0.7295. It had a good generalization and little overfitting because of its close similarity between training and testing. Dam quality, climate change, topography and drainage, siltation, monsoon intensity, deforestation and river management were found to be the top 7 predictors for the feature importance analysis. The results show that the flood probability is a result of the cumulative effect of environmental pressures, infrastructure condition and socioeconomic preparedness. The proposed framework used in conducting flood-risk screening, prioritizing infrastructure, disaster preparedness, and evidence-based planning. Real-time meteorological, remote-sensing, hydrological, and geospatial data should be included in future studies for temporal forecasting and external generalizability.

Keywords: CatBoost; disaster risk reduction; flood probability; infrastructure resilience; machine learning

1. Introduction

Floods are one of the most destructive natural hazards globally, resulting in significant loss of life, property, ecosystems and national economies. The frequency of flooding has increased in both developed and developing nations due to enhanced urbanization, high rates of land use change, climate variability and extreme precipitation events. Environmental degradation and poor infrastructure planning have compounded the flood disasters that have occurred over the past few decades, making the need for an effective prediction system that can aid disaster preparedness and sustainable water resource management (WRM) more urgent (Chen et al. 2026). Recent global assessments have underscored the need to shift the focus from hazard identification to flood probability and risk estimation in order to take timely action and minimize disaster impacts (Zhai et al. 2026).

The impact of flood disasters has tremendous economic, environmental and social effects. Economic damages involve losses to transportation, residential buildings, farmland, and public facilities, and environmental damages are associated with soil erosion, wetland loss, sediment transport, and water quality impairment. Social impacts range from displacement of people and disruption of livelihoods to public health crises and vulnerability of marginalised communities. These multidimensional effects highlight that a multi-disciplinary approach should be adopted in flood risk management, integrating hydrological, environmental, climatic and socioeconomic aspects (El-Bagoury and Gad 2024). Thus, flood risk assessment has emerged as a crucial component in DRR as timely forecasts allow for the planning of evacuations, protection of infrastructure, response to disasters and long-term adaptation to climate change (El Baida et al. 2024).

The occurrence of flooding is a combination of environmental conditions and infrastructure quality. The effects of climate change have led to more rainfall, climate variability, and extreme weather events that have increased the frequency and intensity of floods around the world (Pasula and Subramani 2025). The intensity of the monsoons is also crucial in the tropics and subtropics where the long wet season can cause flooding when the watershed water supply is overstretched (Fang et al. 2022; Sit et al. 2020).

Other factors that influence flood risk include topography, drainage density, watershed structure, and land conditions. Steep slopes, poor drainage, deforestation, loss of wetlands, and urbanization contribute to decreased infiltration and increased flood susceptibility (Sabzi & Ahmadi, 2025). A lack of strong drainage networks, aging dams, river encroachment and poor urban planning also reduce flood resilience (Darabi et al. 2019). When these environmental and infrastructure factors are introduced, the overall context of flood dynamics is better understood than by using meteorological data alone (Gupta, Kanungo, and Das 2025).

The traditional approaches such as statistical and hydrological models involve high level of assumptions, extensive calibration, and high computational effort, and are not always able to handle the nonlinear relationships (Mosavi, Ozturk, and Chau 2018). However, machine learning (ML) can overcome these limitations by learning from large data sets without making any assumptions in advance (Mosaffa et al. 2022).

Compared to conventional statistical models, ML models produce more accurate results, are more suitable for high-dimensional data and nonlinear relationships and adapt well to different areas.

Flood mapping and forecasting (FMF) have demonstrated their effectiveness with ensemble methods, deep learning, and hybrid models (Arganis-Juárez and Preciado 2026). In recent years, a combination of remote sensing, climate and geospatial data has enhanced prediction and disaster management (Loke, Kho, and Raghunandan 2026). Incorporating spatial and temporal dependencies is achieved by using deep learning models like CNNs and LSTMs, which further improve flood prediction (Feng et al. 2026).

Although much work has been done on flood prediction research, current studies often focus on hydrological and climatic parameters and offer relatively little coverage of infrastructure-related parameters, including drainage systems, river management, dam quality, urban planning, and public infrastructure deterioration. This fragmented approach excludes a holistic evaluation of flood risk because environmental processes and infrastructure conditions interact to affect the occurrence and severity of floods (Heo and Lee 2025). Furthermore, most of the previous studies have used relatively small datasets or geographic regions, limiting the predictive model's applicability to a specific context (Zhai et al. 2026).

To overcome these restrictions, the current study set out to create valuable machine learning models to estimate the probability of flooding, using a comprehensive dataset that combines environmental, climatic, and infrastructure-based indicators. In particular, the study aims to test the predictive ability of several machine learning models and to determine which variables are most important for the prediction of floods. The study aims to offer evidence-based insights to support flood preparedness, disaster risk reduction, infrastructure planning and sustainable environmental management by incorporating diverse environmental and infrastructure factors within the study's unifying predictive framework.

2. Materials and Methods

2.1 Dataset Description

The dataset for flood predictions' main training file had 1117957 observations and 22 columns (one of which was an identifier variable, the remaining 20 being the predictor variables, with the continuous outcome variable being FloodProbability). The test set consisted of 745,305 observations and 20 predictors, but there was no target variable (provided with the accompanying test file) (Khalid 2025). The identifier column was not used in building the model because it did not have any analytical relevance.

The predictors were categorized into three groups: environmental, infrastructure and socioeconomic. Environmental indicators comprised the intensity of monsoon, climate change, topography and drainage, coastal vulnerability, watersheds, loss of wetlands, deforestation, landslides and siltation. River management, dam quality, drainage, deterioration of infrastructure, and poor planning were the infrastructure indicators. Socio-economic indicators such as urbanization, encroachments, population score, agricultural activities, political factors and poor disaster preparedness were identified. In this study, Flood Probability was considered a continuous dependent variable and thus the study was conducted as a supervised regression analysis.

2.2 Data Pre-processing

Data was checked for missing data, duplicate data, mismatched data types, and invalid data. There were no missing values or exact duplicate observations found. The id variable was dropped before analysis and all 20 substantive predictors were kept as they were theoretically relevant to flood risk dimensions.

The predictor variables were numerical and were kept on their original scale. In the primary analysis, no feature scaling was done as the algorithms chosen for the tree-based approaches were not sensitive to differences in the measurement scale. A fixed random seed was used to randomly partition the labelled dataset into 80% training data and 20% holdout testing data. The labelled training dataset contained observed FloodProbability values that were used for model development and performance evaluation. The accompanying test dataset did not include the target variable and was intended only for generating predictions.

2.3 Exploratory Data Analysis

Exploratory analysis was performed to gain insights into the distribution and relationships between the variables of this study. Descriptive statistics of the mean, median, standard deviation, minimum and maximum were provided for all the predictors and the target variable. Distributional patterns, skewness, dispersion and possible extreme observations were investigated using histograms and boxplots. The linear relationships between the predictor variables and FloodProbability were measured using Pearson correlation coefficients. A heatmap was created using the correlation matrix. Exploratory interpretation was conducted using correlation analysis and not for automatic exclusion of features.

2.4 Machine Learning Model Development

It is written that 6 algorithms have been developed for regression problems, which were compared: Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, Extreme Gradient Boosting, Light Gradient Boosting Machine, Cat Boost Regressor. The Decision Tree model was adopted as a baseline while boosting-based models and Random Forest were chosen for their ability to capture non-linear relationships and complex interactions between environmental, infrastructure and socioeconomic indicators.

Hyperparameters were optimized using GridSearchCV with five-fold cross-validation. The optimal parameter combinations identified during model tuning were subsequently used for final model training.

2.5 Model Evaluation

The predictive performance of the machine-learning models was evaluated using mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and the coefficient of determination (R^2). Lower MAE, MSE, and RMSE values indicated better predictive accuracy, whereas a higher R^2 value indicated that the model explained a greater proportion of the variance in flood probability. Five-fold cross-validation was performed on the training dataset to assess model stability and generalizability. The final model was selected based on the lowest holdout MAE, MSE, and RMSE values, the highest holdout R^2 value, consistent performance across the cross-validation folds, and a minimal difference between training and testing performance.

Residual distributions and observed-versus-predicted plots were also examined to identify systematic prediction errors, model bias, and possible overfitting.

2.6 Feature Importance Analysis

To determine the most significant variables for flood probability prediction, feature importance was assessed. The 20 predictors were ranked by the first calculation of the importance scores under the random forest model. The best-performing model was then used with the help of Shapley Additive exPlanations for a more interpretable analysis of feature contribution. A global ranking of predictors was achieved using mean absolute SHAP values. To investigate the extent and direction of effects related to the most influential environmental, infrastructure and socioeconomic indicators, SHAP summary and dependence plots were created.

2.7 Statistical Analysis

Descriptive statistics, Pearson correlation analysis, model performance comparison, cross-validation, residual analysis, and feature importance were all used for statistical analysis. The study used data-analysis and machine-learning libraries implemented in Python for all analyses. Data partitioning and model development were conducted with a fixed random seed for the purposes of ensuring model reproducibility. Findings were assessed based on their predictive accuracy, model stability, the magnitude of their errors, and the contribution of the indicators of flood risk to the substantive model.

3. Results

3.1 Descriptive Statistics of Flood-Risk Indicators

This section presents descriptive statistics of the flood-risk indicators. This also describes the flood-risk indicators. The distributions of the predictor variables were similar across the environmental, infrastructure and socioeconomic domains. All indicators had mean values of around 4.91-4.96, with a median value of 5.00. The range of the standard deviations was around 2.0, which was quite variable among observations. Minimums were mostly equal to zero, and maximums ranged from 15 to 19. No single predictor dominated the analysis because of its range of measurement, as the similarity in scale and dispersion found suggested. The descriptive statistics of the predictor variables for the flood risk prediction model (environmental, infrastructure, and socioeconomic) are summarized in Table 1.

Table 1. Descriptive Statistics of Predictor Variables

Predictor	Domain	Mean	Median	SD	Minimum	Maximum
Monsoon intensity	Environmental	4.92	5.00	2.06	0	16
Topography and drainage	Environmental	4.93	5.00	2.09	0	18
River management	Infrastructure	4.96	5.00	2.07	0	16
Deforestation	Environmental	4.94	5.00	2.06	0	17
Urbanization	Socioeconomic	4.94	5.00	2.08	0	17
Climate change	Environmental	4.93	5.00	2.06	0	17
Dam quality	Infrastructure	4.96	5.00	2.09	0	16

Siltation	Environmental	4.93	5.00	2.07	0	16
Agricultural practices	Socioeconomic	4.94	5.00	2.07	0	16
Encroachments	Socioeconomic	4.95	5.00	2.08	0	18
Ineffective disaster preparedness	Socioeconomic	4.95	5.00	2.08	0	16
Drainage systems	Infrastructure	4.95	5.00	2.07	0	17
Coastal vulnerability	Environmental	4.95	5.00	2.08	0	17
Landslides	Environmental	4.93	5.00	2.08	0	16
Watersheds	Environmental	4.93	5.00	2.08	0	16
Deteriorating infrastructure	Infrastructure	4.93	5.00	2.07	0	17
Population score	Socioeconomic	4.93	5.00	2.07	0	18
Wetland loss	Environmental	4.95	5.00	2.07	0	19
Inadequate planning	Infrastructure	4.94	5.00	2.08	0	16
Political factors	Socioeconomic	4.94	5.00	2.09	0	16

3.2 Correlation Among Flood-Risk Indicators

The 20 predictor variables had generally weak correlation with each other and were found to have little multicollinearity by a Pearson correlation analysis. Each predictor, on the other hand, had a positive correlation with flood probability. None of the indicators showed a linear association so strong as to dominate, indicating flood probability was calculated as a function of multiple variables from the environmental, infrastructure, and socioeconomic state. Figure 1 presents the Pearson correlation matrix for the ten predictors exhibiting the strongest absolute correlations with flood probability.

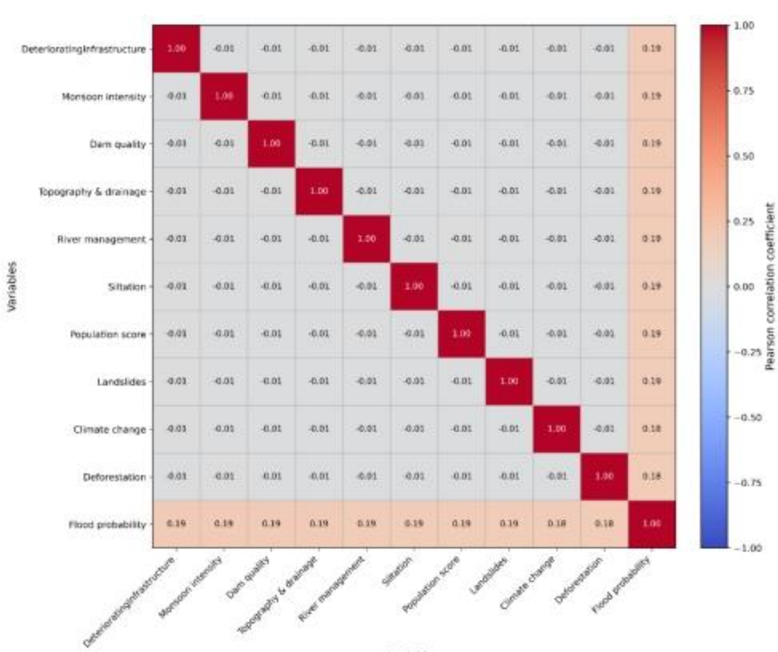

Figure 1. Pearson correlation heatmap of the ten predictors most strongly associated with flood probability in the study dataset

The correlation heatmap demonstrates that the predictors contributed relatively independently to the target variable, supporting their simultaneous inclusion in the machine-learning models.

3.3 Performance Comparison of Machine-Learning Models

There were significant differences in predictive accuracy among the six regression models. The CatBoost model had the fewest errors in the test set, with 0.0220 for MAE, 0.000713 for MSE, and 0.0267 for RMSE. It also had the maximum testing R^2 value of 0.7295. XGBoost and LightGBM are the second and third, respectively, with Random Forest coming in behind. Of all the Decision Trees, the single DT gave the poorest generalization result with an RMSE of 0.0460 and R2 of 0.1959. Table 2 shows the performance of the six machine learning models in terms of MAE, MSE, RMSE and R^2 to determine the most accurate regression model.

Table 2. Comparative Performance of Machine-Learning Models

Model	**Test MAE**	**Test MSE**	**Test RMSE**	**Test R^2**
CatBoost	0.022034	0.000713	0.026702	0.729511
XGBoost	0.024523	0.000884	0.029724	0.664821
LightGBM	0.024875	0.000913	0.030214	0.653680
Random Forest	0.028180	0.001179	0.034341	0.552610
Gradient Boosting	0.031811	0.001512	0.038883	0.426440
Decision Tree	0.036927	0.002120	0.046040	0.195889

3.4 Best-Performing Machine-Learning Model

Among the six evaluated regression algorithms, CatBoost achieved the strongest predictive performance. It produced a training R^2 value of 0.7314 and a holdout-testing R^2 value of 0.7295. Similarly, the training MAE of 0.0217 was close to the testing MAE of 0.0220. The small differences between the training and testing results suggest that the fitted CatBoost model generalized adequately to the holdout data and exhibited limited evidence of overfitting. Table 3 presents the training and testing performance of the CatBoost model.

Table 3. Training and Testing Performance of the CatBoost Model

Dataset	MAE	MSE	RMSE	R^2
Training data	0.015525	0.000379	0.019459	0.853934
Holdout testing data	0.016212	0.000411	0.020271	0.844115

The stability of CatBoost across the training and testing datasets indicated that its superior performance was not attributable solely to fitting the training observations.

3.5 Feature Importance Analysis

The feature-importance analysis showed that many variables were important to flood probability and that no single variable dominated the set. Damage of dam was given the highest Normalized Mean importance score followed by climate change, topography drainage, siltation and intensity of monsoons. There was also a significant contribution from deforestation and river management. The small span of importance values indicated that the target variable contained a balanced set of information from the various environmental, infrastructure, and socioeconomic indicators. The relative importance of the three types of indicators (environmental, infrastructure, and socioeconomic) for predicting flood probability based on the best-performing machine learning model is shown in Figure 2.

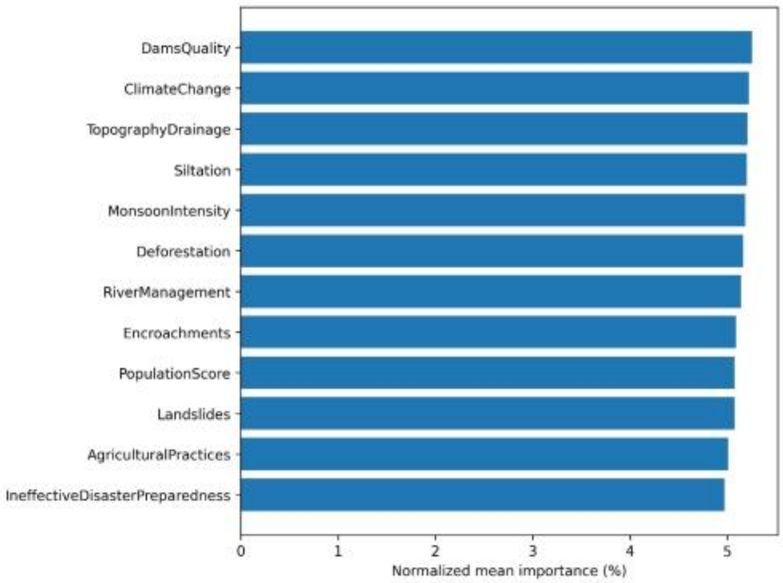

Figure 2. Feature importance ranking of the most influential environmental and infrastructure-based predictors contributing to flood probability estimation

The feature-importance plot shows the top predictors using the normalized importance scores of CatBoost and Random Forest. The feature importance analysis for the best-performing model was carried out and the most influential predictors contributing to the estimation of flood probability are ranked in Table 4.

Table 4. Leading Predictors of Flood Probability

Rank	Predictor	Domain	Mean importance (%)
1	Dam quality	Infrastructure	5.254
2	Climate change	Environmental	5.221
3	Topography and drainage	Environmental	5.205
4	Siltation	Environmental	5.198
5	Monsoon intensity	Environmental	5.180
6	Deforestation	Environmental	5.160
7	River management	Infrastructure	5.139

8	Encroachments	Socioeconomic	5.087
9	Population score	Socioeconomic	5.074
10	Landslides	Environmental	5.073

3.6 Prediction Accuracy

The plot of observed and predicted values revealed that CatBoost predictions were mostly clustered around the 45-degree reference line. This pattern showed good agreement between the estimated and observed flood probabilities over the majority of the outcome range. The dispersion of predictions has also slightly increased in the vicinity of the lowest and highest observed probabilities, suggesting that it is more difficult to predict extreme cases than values near the centre of the distribution. The flood probability values obtained from CatBoost regression model are compared with the observed values in Figure 3 which shows the ability of the model to predict.

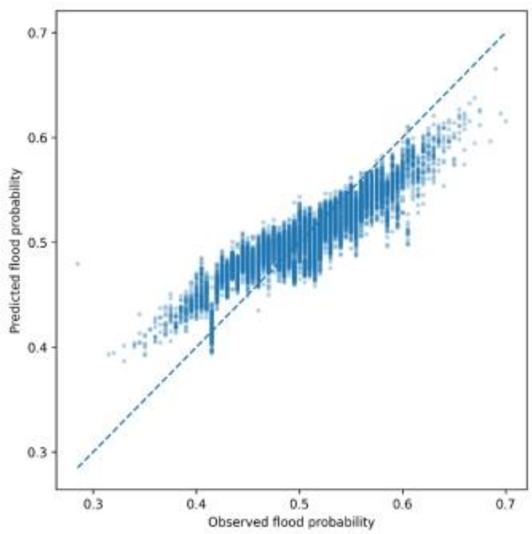

Figure 3. Scatter plot comparing observed and predicted flood probability values for the best-performing CatBoost regression model

The close concentration of observations around the diagonal reference line confirms the comparatively strong predictive accuracy of the CatBoost model.

3.7 Model Error Distribution

Residual analysis further supported the adequacy of the selected model. The mean residual was approximately zero, indicating an absence of systematic overall underprediction or overprediction. The residual standard deviation was 0.0267, while the median residual was 0.0006. Approximately 90% of residuals occurred between −0.0434 and 0.0401. Although a limited number of larger negative and positive errors were observed, the residual distribution remained strongly centred around zero. Figure 4 shows the distribution of residual errors for the CatBoost model, providing an assessment of prediction error and overall model fit.

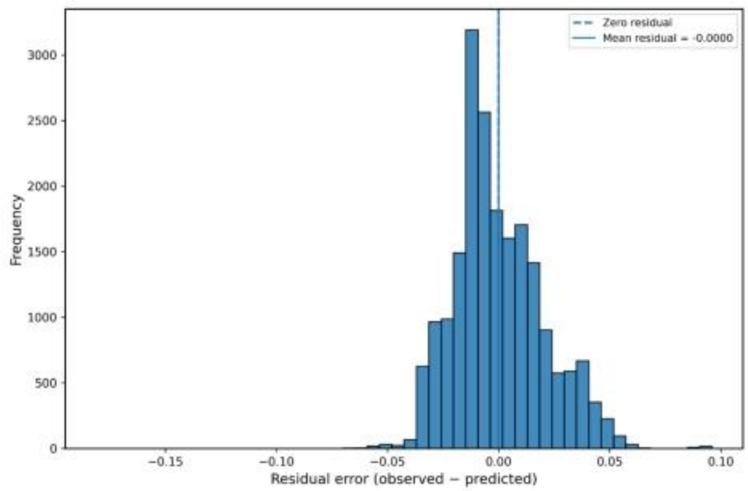

Figure 4. Distribution of residual errors for the CatBoost model showing prediction accuracy and overall model fit

The concentration of residuals around zero and their approximately symmetric distribution suggest that the CatBoost model did not exhibit substantial systematic overprediction or underprediction in the holdout dataset.
Overall, the findings indicated that the evaluated boosting models outperformed the individual Decision Tree and Random Forest models within the supplied dataset.

4. Discussion

The present study involved creating and comparing six machine learning regression models for flood risk prediction based on environmental, infrastructure, and socioeconomic data. The results of all the algorithms evaluated showed that CatBoost outperformed the others, with XGBoost and LightGBM following, highlighting the value of ensemble boosting algorithms in capturing intricate flood-risk relationships. The feature importance analysis also showed that flood probability is not dictated by any single predictor but rather by a combination of climatic factors, terrain attributes, infrastructure quality and human activities. The results confirm the recent evidence that advanced machine learning methods possess the capability to model nonlinear interactions between multiple flood-driving variables to offer strong predictive power for flood susceptibility assessment and hydrological modelling (Perera et al. 2026; Santos et al. 2025).
Environmental variables were among the most influential predictors of flood probability. Climate change, monsoon intensity, topography and drainage, siltation, loss of wetlands and deforestation were always among the most important predictors with direct impacts on runoff generation, infiltration capacity and watershed response. The results are in line with earlier studies indicating that hydrological processes are highly sensitive to alterations in climatic conditions and landscape attributes especially in areas with fast-changing landscapes (Sit et al., 2020). In the same way, optimized tree-based machine learning models have proven that incorporating multiple environmental variables significantly enhances flood susceptibility prediction over traditional statistical methods (Eslaminezhad et al. 2022). Landslide susceptibility is the main theme of the study, but recent investigations have highlighted that the common environmental triggers for

multiple hydrogeomorphological hazards are climate variability, terrain attributes and land-cover changes (Nhu et al. 2020).

Flood probability was greatly affected by infrastructure variables. Engineered systems proved to be a strong predictor, with aspects such as dam quality, drainage, river management and infrastructure condition all playing a role. Lack of maintenance and inadequate drainage decrease the ability of runoff, raising flood danger. These findings corroborate studies that highlight the need for a holistic approach to flood prediction and urban resilience by considering infrastructure and hydrological elements (Amponsah et al. 2026; Hoang and Liou 2024).

The high accuracy of CatBoost and other boosting models is like previous studies on flood susceptibility. Ensemble and hybrid ML approaches are demonstrated to be superior to traditional approaches to capture nonlinear relationships between the environment (Tehrany, Kumar, and Shabani 2019). In addition, recent studies have validated that ensemble and deep learning models usually give higher accuracy in flood prediction and hazardous mapping (Silwal et al. 2026; Waqas and Nawaz 2026). Advanced AI-based frameworks further support their effectiveness in complex environmental modelling (Al-Rawas et al. 2024).

The suggested ML framework is a useful approach for flood risk assessment and disaster plans. It combines environmental, infrastructure and socioeconomic information and is used to identify vulnerable areas, to direct investments in infrastructure and to aid in emergency planning. These types of models based on data are particularly helpful in rapidly urbanizing areas where land use is constantly evolving. The same methods are useful for sustainable planning and disaster management (Zivkovic et al. 2021) and the integration of land-use and climate information enhances hazard forecasting in changing environments (Viet Du et al. 2023).

A major strength of the present study is that it relies on a vast labelled dataset with more than one million observations and twenty flood-related indicators spanning environmental, infrastructure and socio-economic aspects. A comparative evaluation of six state-of-the-art machine learning algorithms gave a thorough evaluation of the predictive performance and reduced algorithm-selection bias. These data consisted mainly of indicators in a structured numerical format, without using high-resolution remote sensing data, real-time meteorological data, river discharge data or temporal rainfall sequences.

In the future, work should be done to incorporate satellite observations, IoT sensor networks, radar rainfall products, and high-resolution DEMs to enhance the prediction accuracy in space and time. Future advancements in model transparency and predictive ability could be achieved through the use of explainable artificial intelligence (XAI) techniques, graph neural networks, transformer architectures, and physics-informed deep learning models. Multi-hazard modelling that jointly considers floods, landslides, and associated environmental hazards may also help to build a more comprehensive framework for climate-resilient disaster risk management and sustainable planning of infrastructure.

5. Conclusion

The current study tested the effectiveness of six machine-learning regression algorithms to predict the FloodProbability outcome based on 20 environmental, infrastructure and socioeconomic indicators. CatBoost was the most successful algorithm, as its prediction errors were the lowest, and the R2 value found was the highest when compared to other algorithms evaluated. The training and testing accuracy were quite similar, suggesting that it can generalise to the provided data. Feature importance analysis also revealed several features such as dam quality, climate change, topography and drainage, siltation, monsoon intensity, deforestation and river management as the features influencing the predicted outcome. The results illustrate the feasibility of ensemble boosting models to model interrelated flood-related indicators among structured indicators. However, there was no temporal, geographic, meteorological, hydrological or remote sensing data in the dataset. Hence, the proposed model must be considered as a predictive tool for the data set of numbers provided rather than as a system for forecasting floods or flood risk at a specific location. The framework must be validated with real-world data from other regions and time periods before it can be used for disaster preparedness, infrastructure prioritization or policy decisions. Future studies could combine the data of rainfall, river discharge, water levels, satellite observation, digital elevation model, land-use and IoT sensor readings to enhance the spatial validity, temporal forecasting capacity, interpretability, and applicability.

References

1. Al-Rawas, Ghazi, Mohammad Reza Nikoo, Malik Al-Wardy, and Talal Etri. 2024. “A Critical Review of Emerging Technologies for Flash Flood Prediction: Examining Artificial Intelligence, Machine Learning, Internet of Things, Cloud Computing, and Robotics Techniques.” Water 16(14):2069.
2. Amponsah, Christopher Atta, Barnali Dixon, Yi Qiang, and Eugene Y. Yan. 2026. “The Application of Geospatially-Integrated Machine Learning Models for Flood Prediction: A Review.” Journal of Hydrology 136016.
3. Arganis-Juárez, Maritza Liliana, and M. Preciado. 2026. “A Methodological Review of AI/ML and Evolutionary Computation in Hydrology and Hydraulics: From ANN/SVM to Deep Learning, Symbolic Regression, and Physics-Informed Models.” Water Science & Technology wst2026304.
4. Chen, Zhenghang, Meili Feng, Matthew F. Johnson, Nigel Wright, Ying Weng, Faith Ka Shun Chan, and Feng Wu. 2026. “Flash Floods in Mountainous Regions: Global Research Trends, Process Mechanisms, and Control Measures.” EGUsphere 2026:1–42.
5. Darabi, Hamid, Bahram Choubin, Omid Rahmati, Ali Torabi Haghighi, Biswajeet Pradhan, and Bjørn Kløve. 2019. “Urban Flood Risk Mapping Using the GARP and QUEST Models: A Comparative Study of Machine Learning Techniques.” Journal of Hydrology 569:142–54.
6. El Baida, Maelaynayn, Mohamed Hosni, Farid Boushaba, and Mimoun Chourak. 2024. “A Systematic Literature Review on Classification Machine Learning for Urban Flood Hazard Mapping.” Water Resources Management 38(15):5823–64. doi:10.1007/s11269-024-03940-7.
7. El-Bagoury, Heba, and Ahmed Gad. 2024. “Integrated Hydrological Modeling for Watershed Analysis, Flood Prediction, and Mitigation Using Meteorological and Morphometric Data, SCS-CN, HEC-HMS/RAS, and QGIS.” Water 16(2):356.
8. Eslaminezhad, Seyed Ahmad, Mobin Eftekhari, Aliasghar Azma, Ramin Kiyanfar, and Mohammad Akbari. 2022. “Assessment of Flood Susceptibility Prediction Based on

Optimized Tree-Based Machine Learning Models." Journal of Water and Climate Change 13(6):2353–85.

9. Fang, Lei, Jinliang Huang, Juntao Cai, and Vilas Nitivattananon. 2022. "Hybrid Approach for Flood Susceptibility Assessment in a Flood-Prone Mountainous Catchment in China." Journal of Hydrology 612:128091.
10. Feng, Xingyu, Kairong Lin, Luwen Zhuang, Yuanhao Xu, and Tongfang Li. 2026. "A Rapid Flood Inundation Integrated Surrogate Model Based on a CNN-LSTM-ATT Deep Learning Framework." Journal of Hydrology: Regional Studies 64:103166.
11. Gupta, Neha, D. P. Kanungo, and Josodhir Das. 2025. "Multi-Hazard Susceptibility Mapping of Landslides and Earthquakes in Bhagirathi Valley Region of Uttarakhand Himalaya, India." Journal of Spatial Science 70(3):403–28. doi:10.1080/14498596.2024.2409083.
12. Heo, Sujung, and Dong Kun Lee. 2025. "Assessing Disasters in East Kalimantan: Machine Learning Approaches for Sustainable Urban Development." Environmental Research Letters 20(10):104003.
13. Hoang, Duc-Vinh, and Yuei-An Liou. 2024. "Elevating Flash Flood Prediction Accuracy: A Synergistic Approach with PSO and GA Optimization." Natural Hazards and Earth System Sciences Discussions 2024:1–26.
14. Khalid, Naiya. 2025. "Flood Prediction Dataset." https://www.kaggle.com/datasets/naiyakhalid/flood-prediction-dataset.
15. Loke, Wyn Shern, Jun Cheng Kho, and Mavinakere Eshwaraiah Raghunandan. 2026. "Artificial Intelligence for Remote Sensing-Based Detection and Prediction of Landslides in Malaysia and Vietnam: A State-of-the-Art Review." Natural Hazards 122(14):547. doi:10.1007/s11069-026-08316-1.
16. Mosaffa, Hamidreza, Mojtaba Sadeghi, Iman Mallakpour, Mojtaba Naghdyzadegan Jahromi, and Hamid Reza Pourghasemi. 2022. "Application of Machine Learning Algorithms in Hydrology." Pp. 585–91 in Computers in earth and environmental sciences. Elsevier.
17. Mosavi, Amir, Pinar Ozturk, and Kwok-wing Chau. 2018. "Flood Prediction Using Machine Learning Models: Literature Review." Water 10(11):1536.
18. Nhu, Viet-Ha, Ayub Mohammadi, Himan Shahabi, Baharin Bin Ahmad, Nadhir Al-Ansari, Ataollah Shirzadi, John J. Clague, Abolfazl Jaafari, Wei Chen, and Hoang Nguyen. 2020. "Landslide Susceptibility Mapping Using Machine Learning Algorithms and Remote Sensing Data in a Tropical Environment." International Journal of Environmental Research and Public Health 17(14):4933.
19. Pasula, Abhishek, and Deepak N. Subramani. 2025. "Data Driven Deep Learning for Correcting Global Climate Model Projections of SST and DSL in the Bay of Bengal."
20. Perera, Manimeldura Devki Devanga, Athanasios Angeloudis, Adil Siripatana, and Lindsay Beevers. 2026. "Hybrid Methods in Flood Inundation Modeling: A Systematic Review." Natural Hazards 122(8):357. doi:10.1007/s11069-026-08078-w.
21. Santos, Leonardo B. L., Elton V. Escobar-Silva, Luiz F. Satolo, Ricardo S. Oyarzabal, Michael M. Diniz, Rogério G. Negri, Glauston R. T. Lima, Stephan Stephany, Jaqueline A. J. P. Soares, Johan S. Duque, Fernando L. Saraiva Filho, Angélica N. Caseri, and Luiz Bacelar. 2025. "Machine Learning-Based Hydrological Models for Flash Floods: A Systematic

Literature Review.” Smart Construction and Sustainable Cities 3(1):21. doi:10.1007/s44268-025-00071-9.

22. Silwal, Abinash, Anil Subedi, Rajee Tamrakar, Kshitij Dahal, Dewasis Dahal, Kenneth Okechukwu Ekpetere, and Mohamed Zhran. 2026. “A Comprehensive Review of Machine Learning and Deep Learning Methods for Flood Inundation Mapping.” Earth 7(2):44.
23. Sit, Muhammed, Bekir Z. Demiray, Zhongrun Xiang, Gregory J. Ewing, Yusuf Sermet, and Ibrahim Demir. 2020. “A Comprehensive Review of Deep Learning Applications in Hydrology and Water Resources.” Water Science and Technology 82(12):2635–70.
24. Tehrany, Mahyat Shafapour, Lalit Kumar, and Farzin Shabani. 2019. “A Novel GIS-Based Ensemble Technique for Flood Susceptibility Mapping Using Evidential Belief Function and Support Vector Machine: Brisbane, Australia.” PeerJ 7:e7653.
25. Viet Du, Quan Vu, Huu Duy Nguyen, Viet Thanh Pham, Cao Huan Nguyen, Quoc-Huy Nguyen, Quang-Thanh Bui, Thanh Thuy Doan, Anh Tuan Tran, and Alexandru-Ionut Petrisor. 2023. “Deep Learning to Assess the Effects of Land Use/Land Cover and Climate Change on Landslide Susceptibility in the Tra Khuc River Basin of Vietnam.” Geocarto International 38(1):2172218. doi:10.1080/10106049.2023.2172218.
26. Waqas, Muhammad, and Mohsin Nawaz. 2026. “Machine Learning and Deep Learning in River-Basin Modeling: A Comprehensive Review.” Applied Soft Computing 115295.
27. Zhai, Xiaoyan, Yongyong Zhang, Jun Xia, Yongqiang Zhang, Qiuhong Tang, Quanxi Shao, Junxu Chen, and Fan Zhang. 2026. “Flood Predictions from Metrics to Classes by Multiple Machine Learning Algorithms Coupling with Clustering-Deduced Membership Degree.” Journal of Geographical Sciences 36(1):149–76. doi:10.1007/s11442-026-2442-8.
28. Zivkovic, Miodrag, Nebojsa Bacanin, K. Venkatachalam, Anand Nayyar, Aleksandar Djordjevic, Ivana Strumberger, and Fadi Al-Turjman. 2021. “COVID-19 Cases Prediction by Using Hybrid Machine Learning and Beetle Antennae Search Approach.” Sustainable Cities and Society 66:102669.