Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Interpretable Machine Learning Framework for Flood Risk Prediction: A Comparative Study of Base Classifiers and Voting Ensembles

Md. Sharfuddin

sharfuddin.seu@gmail.com

Southeast University

Md. Shakil

Southeast University

Md Sabbir Hossain Siam

Southeast University

Research Article

Keywords: Flood Risk Prediction, Machine Learning, Logistic Regression, Disaster Management, Environmental Modelling, Predictive Analytics

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11844484/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

An Interpretable Machine Learning Framework for Flood Risk Prediction: A Comparative Study of Base Classifiers and Voting Ensembles

Md. Sharfuddin [1]*, Md Shakil [2] and Md Sabbir Hossain Siam [3]

[1]* Department of Computer Science and Engineering, Southeast University, Dhaka, Bangladesh
sharfuddin.seu@gmail.com

[2] Department of Computer Science and Engineering, Southeast University, Dhaka, Bangladesh
imshakil.me@gmail.com

[3] Department of Computer Science and Engineering, Southeast University, Dhaka, Bangladesh
its.me.ssiam@gmail.com

Correspondence: sharfuddin.seu@gmail.com *

Abstract

Floods are among the most destructive natural hazards and have considerable impacts on human life, infrastructure, agriculture and economic stability worldwide. The rise in the frequency and severity of flood events caused by climate variability and environmental changes has led to a pressing demand for precise and effective flood risk prediction systems. Traditional hydrological models are often based on complex simulations, extensive physical parameters and require a lot of computational resources, which limits their applicability for rapid decision making. Machine learning (ML) techniques offer a promising alternative by extracting complicated patterns from environmental data and producing dependable predictive outcomes. However, existing studies are mostly limited to individual algorithms or computationally intensive models, and little attention has been paid to the comprehensive comparative analysis among classical ML approaches with interpretable frameworks. In this study, we develop a machine learning-based flood risk prediction framework using the Flood Prediction Dataset from Kaggle. Nine machine learning algorithms were systematically evaluated including Logistic Regression (LR), Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), XGBoost, AdaBoost and LightGBM. Also, the study looked into ways of merging features to find the optimal predictive configuration. The experimental results indicate that Logistic Regression has the best individual performance with accuracy of 85.73%, precision of 86.15%, recall of 85.84% and F1-score of 85.99%. The results show that a straightforward interpretable ML model can predict flood risks successfully and with a competitive performance to complex ensemble methods. The proposed framework provides a lightweight and practical solution for supporting the early flood warning and disaster management applications.

Keywords: Flood Risk Prediction; Machine Learning; Logistic Regression; Disaster Management; Environmental Modelling; Predictive Analytics.

I. INTRODUCTION

1.1 Background and Problem Statement

Floods are among the most common and lethal phenomena of disaster in the world, claiming hundreds of thousands of lives each year, displacing millions and causing billions of dollars of damage to infrastructure and economic loss. Floods are among the most common disasters in the world and, according to UNDRR, they are becoming more frequent due to climate change, rapid urbanization and environmental degradation[1] . Flooding is a major public health problem, causing injuries, waterborne diseases, forced displacement and disruption of essential services, the World Health Organisation (WHO) said [2] .

Climate change influenced the frequency of extreme precipitation events, the hydrological cycle, and the susceptibility of densely populated areas. The Intergovernmental Panel on Climate Change (IPCC) said rising global temperatures had been accompanied by changes in the variability of rainfall and extreme weather events, making floods more likely in many areas [3]. Consequently, accurate flood forecasting systems are essential for disaster preparedness and emergency response, as well as for climate adaptation strategies.

Historically, flood forecasting has been based on hydrological and hydraulic models such as rainfall-runoff models and river flow models. These methods provide important physical insights but require many environmental parameters, targeted geographical regions and substantial computational costs. In addition to the availability of good quality input data, the accuracy of such models is also highly dependent on their calibration models can limit their application in many data-constrained regions [4] , [5] .

1.2 Rationale of the Study

Part of this progress is due to the growing availability of environmental data sets, and advances in artificial intelligence, which have opened up new possibilities for improving flood prediction systems. This deep learning approach guarantees that the model learns a complex correlation between environmental factors and the occurrence of floods without explicit physical assumptions. These features emphasize the promising nature of the application of ML-based methods for efficient, scalable and data-driven flood forecasting frameworks.

Now a days Support Vector Machine, Random Forest, Artificial Neural Networks, Gradient Boosting and Logistic Regression are the most used machine learning techniques for flood susceptibility mapping and risk classification. However, most of the existing studies focus on advanced ensemble or deep learning models ignoring the results with simpler interpretable models . In practice, disaster management models have to strike a balance between prediction performance, computability and interpretability.

Therefore, this study aims to overcome the limitation by developing a robust and computationally inexpensive flood risk prediction framework that compares variety of machine learning algorithms and selects an optimal model for application.

1.3 Related Work Review and Limitations

Machine learning methods have been demonstrated to be within the potential of hose, for instance, from previous flood prediction Given the possible superior performance of many ensemble models (Random Forest, XGBoost, Gradient Boosting) in prediction problems with nonlinear relationships among environmental variables. Deep learning models have been able to find complex features in very large datasets and achieved competitive results.But there are a number of limitations.

Comparative evaluation has been limited to:

Most of the studies use one or two machine learning algorithms, which makes it difficult to identify the best model.

Model complexity: high
Deep learning and ensemble methods are complex ones that typically require massive amounts of data and lots of computational power

Interpretability is limited:
Complex models typically operate in a black-box model, making the decision transparent to the decision makers.

Insufficient analysis of combinations of features:
Most of the previous studies did not systematically study the contribution from group of features to the performance of flood prediction.

Challenges in practical deployment:
However, accuracy only trained models might not meet the constraints of real-time disaster management applications.

The limitations highlight the need for a comprehensive evaluation framework to compare multiple ML models, considering accuracy, simplicity and real-world relevance.

1.4 Suggested Approach

The proposed study suggests a machine learning-based flood risk prediction framework using the Flood Prediction Data Set available on Kaggle. This study analyzes nine popular machine learning algorithms:

•Logistic Regression (LR)
• Support vector machine (SVM)
• Naive Bayes (NB)
• K-Nearest Neighbour(KNN)
• Decision Tree (DT)
• Random Forest (RF)
• XGBoost
• AdaBoost
• Light GBM

The models are trained and evaluated on some basic performance indicators such as accuracy, precision, recall, F1-score and mean square error. Finally we do feature combination experiments to see if combining good performing feature groups can improve the prediction power.
The overall Logistic Regression model gave the best individual classifier with an accuracy of 85.73%, this finding further supports that an interpretable statistical learning model can ease the flood risk prediction without a complex architecture.

1.5 Contributions of the study

The major contributions of our study can be summarized as follows:
Development of a new prediction framework of floods from multiclass machine learning results for classification of risk. Evaluation of nine ML models to determine the best prediction approach. Study of feature combination strategies to investigate effect on flood prediction performance. Among the classifiers considered, Logistic Regression was the best performing classifier. Presentation of an interpretable and lightweight prediction method, which has been shown to be more implementable for practical solutions in flood early-warning and disaster management systems.

II. RELATED WORK

Srivastava et al. [6] attempted to develop a flood prediction framework by taking into account the variables from the environmental, hydrological, infrastructure and socio-economic domains of the 50 000-record dataset. The Bison Optimization based Efficient Deep Neural Network has an accuracy of 95.2% which is higher than RF and SVM (Except SVM is only 79.2% due to false positives.) However, the model was computationally expensive and had not been validated against any independent real-world historical flood events.

Cui et al. [7] proposed an improved Kernel Extreme Learning Machine to predict the probability of flood. The results indicated that the MESAO-KELM model achieved the smallest RMSE (4.6674×10−5) and MAE (2.9504×10−5) compared to the SVM, RF, ELM and CNN models. However, only 3000 samples were used and the model was not validated using real time temporal flood data.

Mancy et al. [8] proposed a distributed flood-control framework with the integration of IoT, federated learning, and reinforcement planning decision-making in multi-agent setups. In particular, they obtained a R^2 score of 0.953, mean squared error (MSE) of 0.112 and mean absolute error (MAE) of 0.207 for their FL-MAPPO model. It offers better privacy and response time but the architecture is a bit complex and heavily dependent on stable communication and computing infrastructure.

Umadevi et al. [9] proposed a novel Temporal Convolutional Network for flood and multi-disaster forecasting using distributed IoT data. The model gave RMSE of almost 0.10 and was about 92.16% accurate in predicting occurrence or non-occurrence of flash flooding in the next hour. But the method required a large amount of time for training and has not been tested in a large scale real-life disaster situation.
Irawan et al. [10] converted the probability of flooding into 4 classes of risk and tested several machine learning and deep learning models. C4. C4. 5, MLP, RF and Logistic Regression

achieved accuracy over 94% on the test set (around 1%). 5 top performers. The transformation to categories is also problematic, since it causes a loss of information and the results are also sensitive to this threshold.

Sovia et al. [11] It combines multilayer stacking ensemble, meta-feature augmentation and Optuna based tuning for flood-risk classification. Xgboost as optimal feature alone has achieved 98.69% accuracy. The ensemble works well for high complexity, but is largely dependent on the goodness of the benchmark dataset, limiting its generalizability across regions.

Wang et al. [12] proposed a two-step method with Random Forest feature importance and an MLP model to predict the flood probability in February 2024. The baseline MLP achieved (R^2) of 79.63 and the optimized model improved it to 85.27 with MSE = 0.000368. The study was not validated at basin, time scale or spatial independence.

Zhang et al. [13] used Random Forest feature ranking and combined it with a neural network to predict the probability of flooding based on environmental and infrastructural variables. The best performing method was the RF-guided neural network that down-selected the number of irrelevant predictors. The study, however, did not report specific numerical test results or an independent validation protocol.

Prakash et al. [14] mapped flood inundation and susceptibility in the state Assam, India using Sentinel1 SAR images, rainfall, terrain soil and land-cover data. Total accuracy reported is 82.91%. Random Forest and other algorithms such as SVM and CART are unable to cross the 85% mark. The model was limited to one geographical area and some overestimation of flood was seen.

El-Mahdy et al. [15] also used neural networks and conventional machine learning algorithms for classifying floods in South Sudan APREST. Gradient Boosting 93.7% Feed-forward neural network 95% accuracy. But they were based on data from one country, and not tested to see if they would hold true in other flood-prone places.

H. Hahn et al. [16] We trained and tested nine deep-learning models with 19 years of data from weather and river sensors in the Wupper River SegRNN achieved the overall accuracy of 99.7% and approximately 91% F1 score for flood detection against the ground truthmodell. In the whole year 2015 evaluation, it showed good performance, though many floods were missed at longer forecast horizons and with missing sensor input.

Alam et al. [17] proposed a set of models for flood prediction using ensemble model in Kosi Megafan where they used Bayesian optimized ensemble of RF, XGBoost, Catboost and LSTM. The ensemble correctly classified about 90% cases and got a Cohen's kappa = 0.79. The SHAP showed rainfall, elevation, drainage density and curve number as the most influential predictors. However, the model is able to use a range of uncertainty factors to generate susceptibility rather than exact flood timing or hydrographs.

Singha et al. [18] used Analytic Hierarchy Process and six machine-learning models to generate flood susceptibility maps. The AHP–Gradient Boosting model achieved an area under the curve (AUC) of 0.996. Hybrid models outperform most of the base learners (AUC>0.988). However, the SHAP weighting is based on subjective expertjudgements and multiple

univariate factor mapping (UM class) maps that generate floods are traditionally precipitation records that do not generate real time flood [19] forecasts.

Xu et al. [19] CNN coupled hydrodynamic simulation, attention mechanisms, Particle Swarm Optimization and SHAP for urban flood-depth 5 The proposed PSOAM CNN produced NSE of 0.9503 for the whole NS, whereas only 0.8341 was obtained for the five-year flood case prediction. This model was site specific and incorporated the assumptions and uncertainties of the associated hydrodynamic simulation.

Hossain et al. [20] mapped flood susceptibility of Feni District, Bangladesh using RF and MaxEnt with fourteen geospatial variables. MaxEnt (AUC ≈0.98) and RF (accuracy and F1 score of ∼93%), respectively. But this only produced spatial maps of flood susceptibility, without any prediction of timing, depth or duration.

Chardi et al. [21] evaluated flood susceptibilities in Ourika Watershed, Morocco using XGboost, Random Forest and KNN. Models averaged an Area Under the Curve (AUC) of ∼0.89 and captured the highly vulnerable mountainous regions . But the study was made from only 557 flood and non-flood observations from one watershed.

Kafi et al. [22] Developed explainable AI to estimate the urban flood risk in Bauchi using applied random forest, XGBoost and SVM. Random Forest AUC was 0.93 whereas SVM AUC was 0.84 with precision of 75.7%. The above study improves our understanding of flood risk factors, but does not provide prediction on flood timing, duration and inundation depth.

Wahba et al. [23] compared ANN-MLP with SVR and other machine learning methods for flood-susceptibility mapping based on 224 flood and non-flood locations The AUC of the SVR model was 95.83 %. After the elimination of variables, the optimized ANN-MLP had an AUC value of approximately 96.7%. However, the dataset was small and from only one study site.

Yosri et al. [24] RAPFLO is a deep-learning framework proposed to predict the flood-risk level, extent and future likelihood across Calgary. The model was able to predict if flooding occurred at 0.96 R^2, however, the risk-level classification was at best 80% accurate. Performance may also be affected by uncertainty about future climate scenarios and urban development patterns.

Chowdhury et al. [25] Mapping flash-flood susceptibility in north-eastern Bangladesh with ANN, RNN, RFGB and CatBoost considering different rainfall-return periods. ANN showed better performance than the other considered models and was successfully used for the extreme rainfall events. But the study used only 400 labeled points, and did not provide real-time continuous flood forecasts.

Zhao et al. [26] developed a flash-flood warning system based on LSTM for mountainous China based on rainfall, terrain slope and curve number information. It had an 84 percent hit rate and a 9 percent false-alarm rate—much better than any traditional warning. But the accuracy was still below 85%. And the model was not tested on data from outside of China.

Chen et al. [27] The model learned spatial dependencies between monitoring stations and outperformed temporal models that do not explicitly model connectivity between stations. But the results are not published in a big numeric scale and depend on a rather dense & stable monitoring network.

Wu et al. [28] developed a deep learning based surrogate model to effectively predict urban flood depth for rainfall events with different return periods. More interestingly, the model produced a spatial slip pattern similar to the hydraulic simulation, but much faster than the conventional simulation. However, it was trained on simulated data and has poor generalizability outside the rainfall range it was trained on.

Kondo et al. [29] trained the neural network model on the outputs of the hydraulic and hydrological simulations for real-time pluvial flood prediction. The machine-learning model was about 19.8 times faster than the original physical simulation, with similar predictive power. But also had to duplicate drainage and terrain data from specific locations, and auto completed modeling inaccuracies that came from the hydraulic simulator.

A. Mosavi. et al. [30] build A review of flood prediction with machine learning approaches such as . ANN, SVM, decision tree, neurofuzzy systems, evolutionary algorithms and hybrid models. The review shows that individual model performance is often improved by ensemble, decomposition, optimization and hybrid. However, the models reviewed were not directly comparable to each other due to differences in datasets, evaluation metrics and study regions.

III. METHODOLOGY

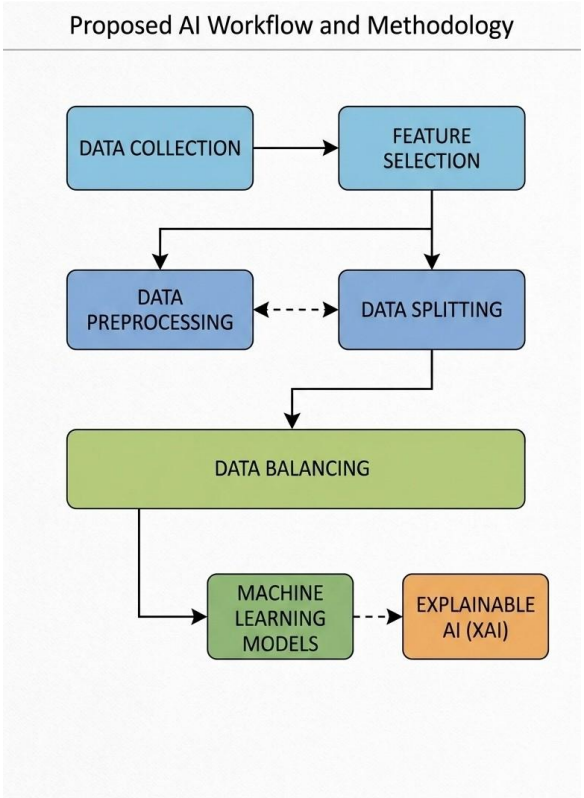

Fig. 1. Workflow for flood risk prediction

A. Data Set and Feature Selection

The dataset for flood prediction was obtained from Kaggle and contains 20 numerical predictors of environmental, infrastructural, climatic and socioeconomic factors and Flood-Probability as the target variable. (Kaggle [20]) The target was classified in binary classes as Flood and No flood. The feature-importance analysis yielded 16 relevant variables.

MonsoonIntensity, PopulationScore, ClimateChange, InadequatePlanning, and PoliticalFactors were the most influential predictors.

B. Data Processing

The data set was checked for missing values, duplicate records and inconsistent observations. Numerical features were standardised, where necessary, to make feature scales comparable and improve the performance of distance- and margin-based classifiers, and the input variables selected were separated from the target.

C. Data Splitting

The processed dataset was divided into training, validation and testing sets with a 70:15:15 ratio. The model was learned on the training set, validated and selected on the validation set and the final performance was evaluated on the independent test set.

D. Data Balancing

To address the class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data only to generate synthetic samples of minority class. We did not change the validation and test sets to prevent data leakage and ensure unbiased evaluation.

E. Machine Learning Models

The nine base classifiers used in this study are SVM, Naive Bayes, KNN, Logistic Regression, Decision Tree, Random Forest, XGBoost, AdaBoost and LightGBM and were implemented in a Python-based Jupyter Notebook. The models LR, SVM and XGBoost were used to select models A, B and C based on their accuracy. Four voting ensemble models were built and compared to the base models, namely A+B, B+C, A+C and A+B+C. A= Highest Accuracy B = 2nd Highest Accuracy C= 3rd Highest Accuracy

F. Explainable AI

The influence of features on model predictions was analyzed using SHAP (SHapley Additive exPlanations). The global feature importance is visualized in Fig. 2 as a SHAP summary plot. The most dominant features are MonsoonIntensity, PopulationScore and ClimateChange. Higher feature values (red dots) correspond to higher predicted target values.

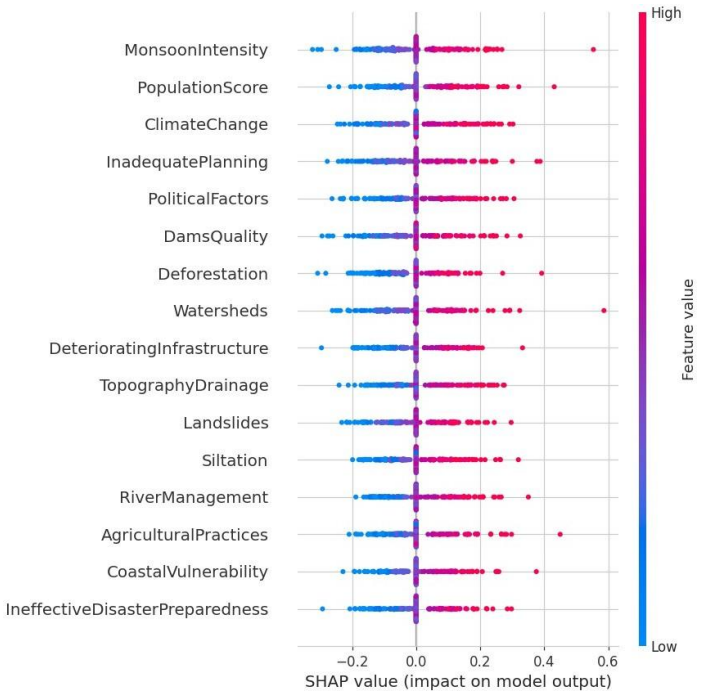

Fig. 2. SHAP Summary Plot

Fig. 3 shows the local interpretability of a single sample using a SHAP waterfall plot. The output begins with a baseline expected value of E[f(X)] = 0.532. The output is increased by positive contributions from Watersheds (+0.11), PoliticalFactors (+0.11) and PopulationScore (+0.10) and decreased by negative contributions from TopographyDrainage (−0.13) and ClimateChange (−0.09) to a final prediction of f(X) = 0.946.

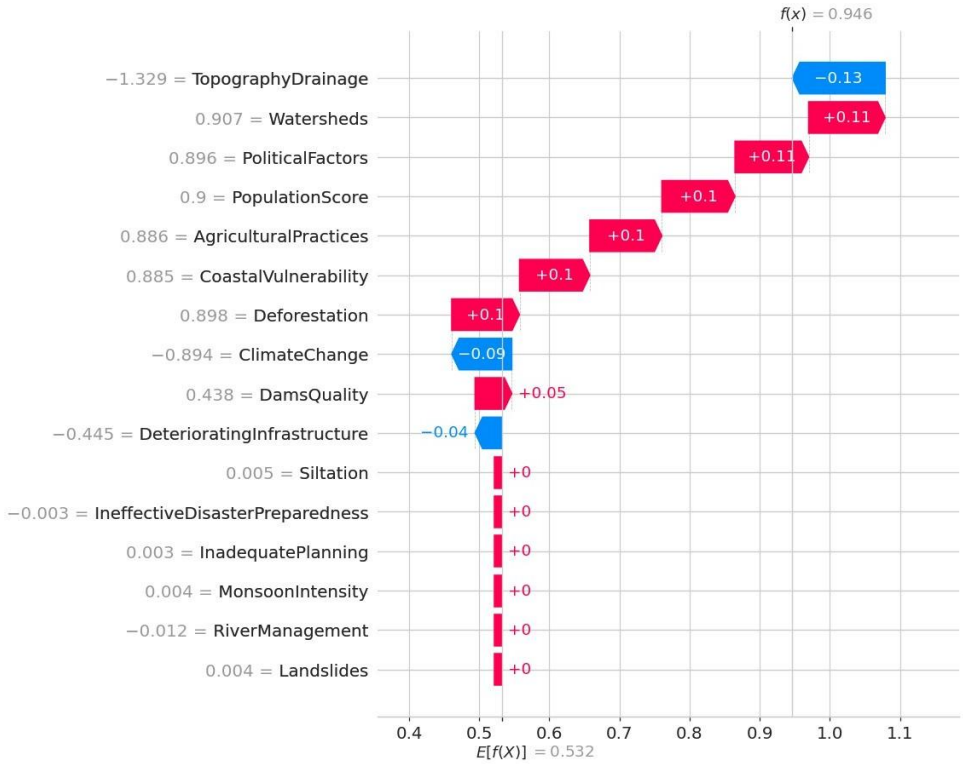

Fig. 3. SHAP waterfall plot

Mean absolute SHAP values over all samples are presented in Fig. 4. MonsoonIntensity has the biggest average impact (0.081), followed by PopulationScore (0.079), ClimateChange (0.075) and InadequatePlanning (0.073). Even lower ranked features such as CoastalVulnerability (0.059) and IneffectiveDisasterPreparedness (0.058) are meaningful in overall performance.

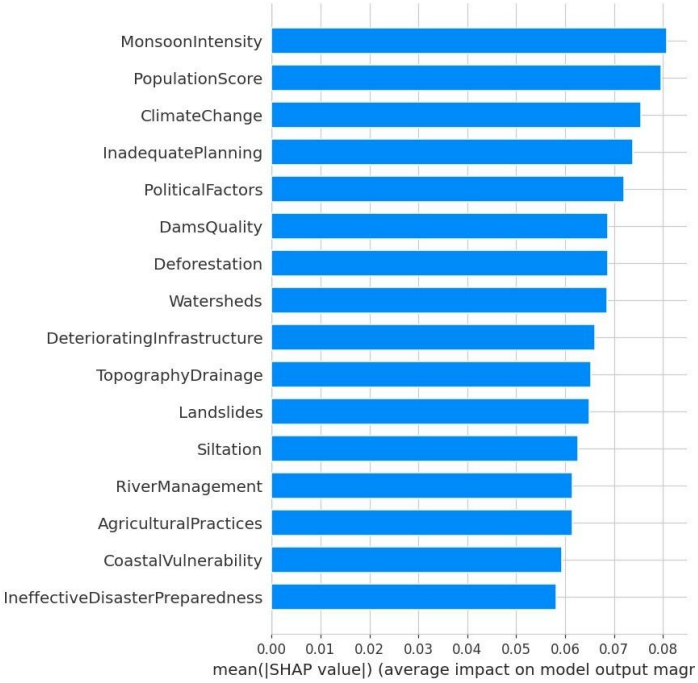

Fig. 4. Mean(SHAP value) (average impact on model output magnitude)

G. Evaluation Metrics

TP (True Positive): Model predicts the area as flood prone and flood occurs at that area. The classifier skill score is the number of flood-prone locations this model correctly identifies.

TN (True Negative): The area is predicted to be not flood prone and it does not flood. In fact, it counts the total number of safe or low risk areas that the model correctly classifies.

FP (False Positive): The model indicates the area is a flooded zone, but there is no flood in that area. It is the number of false flood alerts or the number of sites misdiagnosed as high risk.

FN (False Negative): The model does not predict any flood-prone area but the area actually experiences flooding. Roof, a prediction error which is very relevant to understand your model ability to detect areas affected by floods and consists in quantifying flooded areas that the model has not detected yet.

Accuracy: Accuracy is the ratio of correct predictions to total predictions. It indicates the percentage of total areas correctly classified by your model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision: Precision is the ratio of number of true flood prone areas predicted to total number of areas predicted as flood prone. That gives an idea of how certain the model is when it predicts a high flood risk.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall: How many of the actual flood-prone areas were correctly identified by the model? It indicates the number of flood prone areas correctly identified by a model from ground truth data.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1-score: F1 score is the harmonic mean of precision and recall. Its aim is to quantify the trade-off between missed flooded areas and false flood warnings.

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

MSE (Mean Squared Error): MSE is the average of squared differences between expected and predicted values. It computes the difference (error) between the estimated probabilities of flood risk, flood depths or values associated with continuous variables and a spouse.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5}$$

ROC–AUC: ROC–AUC is the area under the ROC curve, which summarizes the performance of a model in distinguishing between flood-prone and non-flood-prone regions at all computation challenge settings. It detects the ability of model to discriminate between high risk and low risk areas at different classification thresholds.

$$AUC = \int_0^1 TPR(FPR) \cdot d(FPR) \tag{6}$$

IV. RESULT

A. System Specification

The proposed framework for flood-risk prediction was implemented in a Python-based Jupyter Note Book environment. Data preprocessing and numerical analysis were performed using Pandas and NumPy. The model development and evaluation were performed using the Scikit-learn, XGBoost, and LightGBM libraries. The base classifiers that were tested are Support Vector Machine, Naive Bayes, K-Nearest Neighbours, Logistic Regression, Decision Tree, Random Forest, XGBoost, AdaBoost and LightGBM. Moreover, voting-ensemble configurations A+B, B+C, A+C and A+B+C were created to check the advantages of combining multiple classifiers. The model performance was evaluated based on confusion matrices and ROC – AUC curves using training, validation and testing datasets and graphical visualisation was performed using Matplotlib and Seaborn packages. The specification is limited to the software and experimental environment since details about processor, RAM and operating-system were not given.

B. Confusion Matrix

TABLE I
CONFUSION MATRIX

Model	TP	TN	FP	FN
SVM	3231	3157	516	596
NB	3136	3100	573	691

Model	TP	TN	FP	FN
KNN	2887	3097	576	940
LR	**3285**	3145	528	**542**
DT	2480	2571	1102	1347
RF	3194	3011	662	633
XGBoost	3226	3117	556	601
AdaBoost	3230	3104	569	597
LightGBM	3217	3109	564	610
A+B	3261	**3158**	**515**	566
B+C	3240	3143	530	587
A+C	3253	3151	522	574
A+B+C	3258	**3158**	**515**	569

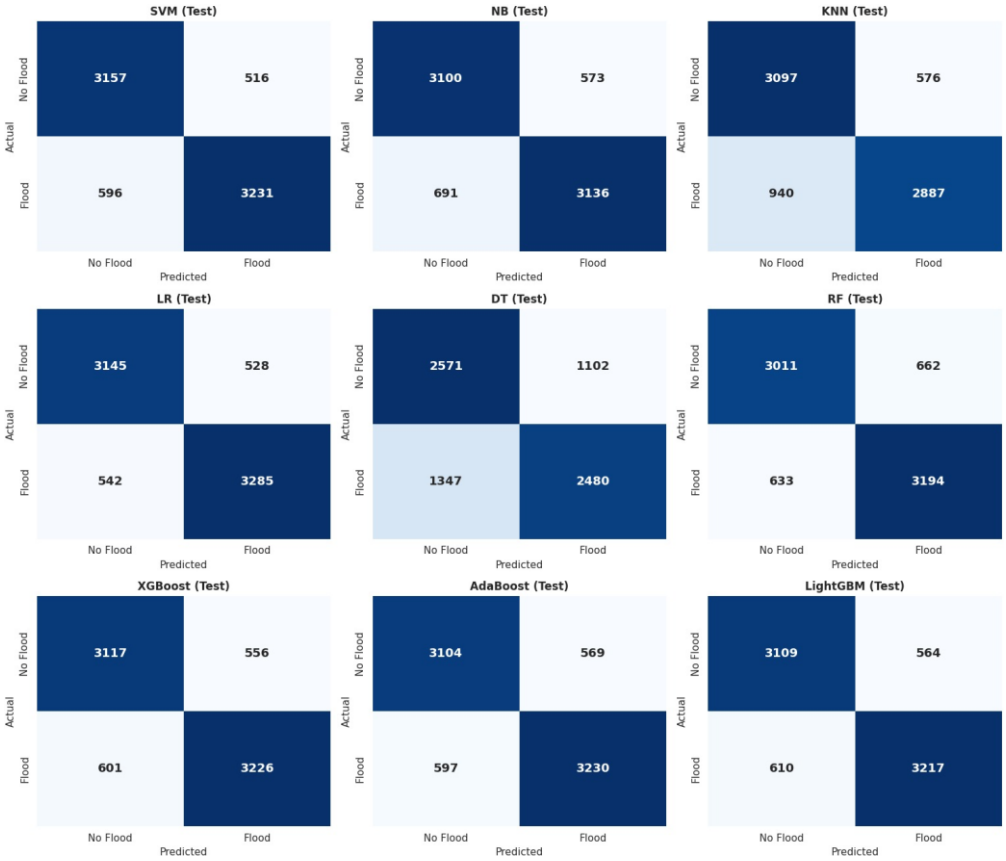

Fig. 5. Confusion Matrix of 9 Base Model

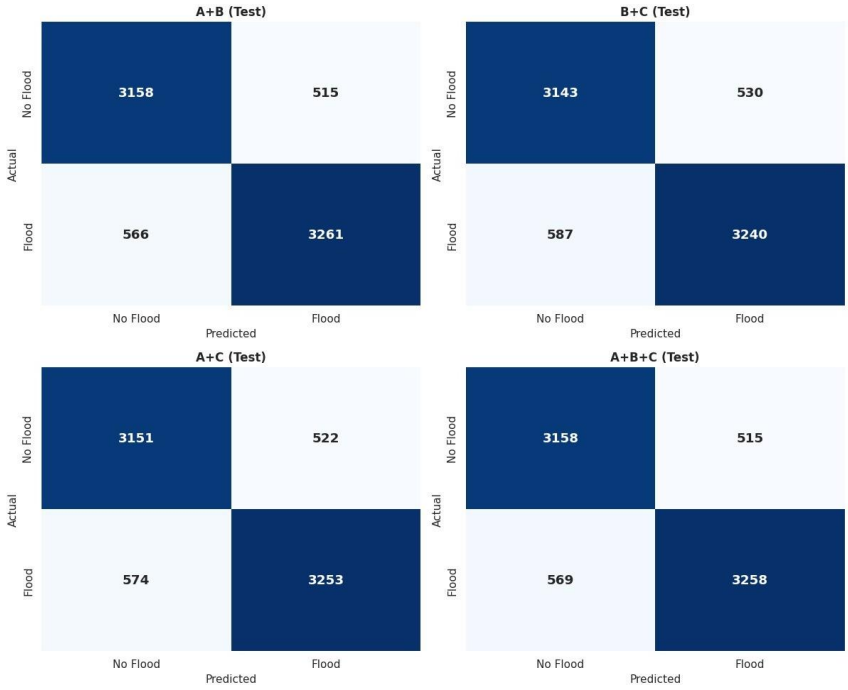

Fig. 6. Confusion Matrix of 4 Ensemble Model

Figure 5: We provide the confusion matrices of the ensemble combinations for flood-risk prediction. The model A+B has the best overall classification performance, A+B+C and A+C are very close, and B+C has a little higher misclassification.

Figure 6: We show the confusion matrices of the single machine learning models . Logistic Regression gives the best flood-detection performance followed by SVM and the boosting-based models while Decision Tree performs the worst in terms of classification capability.

Confusion-matrix results for flood-risk prediction of the machine-learning models evaluated. Logistic Regression (LR) produced the most true positives (3,285), and had the least false negatives (542), suggesting it was best able to identify regions prone to flooding. Thus, as we can see from the above columns that SVM, XGBoost, AdaBoost, LightGBM and Naive Bayes also showed competitive performance while Random Forest and KNN had relatively higher misclassification rates. The overall contribution of ensemble combinations, A+B performed the a good balance with 3,261 true positives, 3,158 true negatives and the lowest false positive (515), followed by A+B+C then A+C and B+C. The Decision Tree (DT) had performance least impact due to having the maximum false-positive values as well as false-negative. In general, LR and the ensemble models came up better in terms of providing more reliable flood-risk classification than individual baseline models.

C. ROC AUC Curve

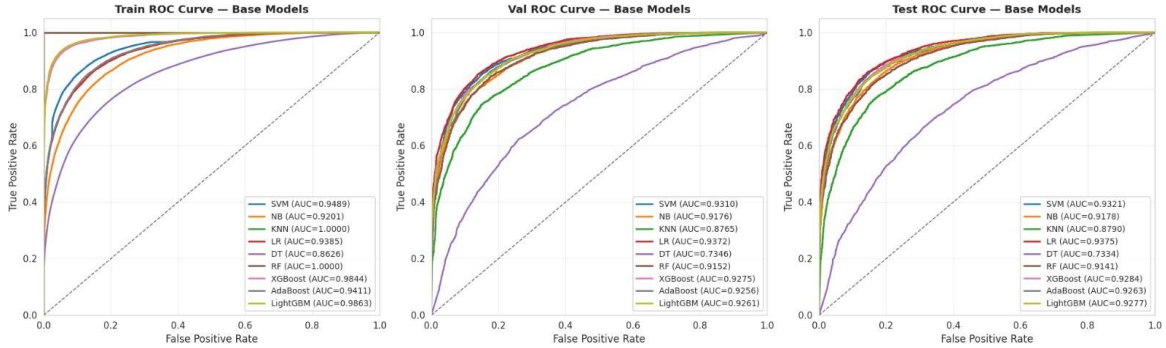

Fig. 7. ROC-AUC Curve (Base Model)

Figure 7: ROC curves of the individual base models for training, validation and test data sets Among the models, Logistic Regression provides the best generalisation summary from validation to test, having the highest AUC on both sets (followed by SVM and finally, boosting based models). The ROC performances from XGBoost, AdaBoost and LightGBM are consistently high in discriminating the two flood-risk classes. The KNN and Random Forest models have perfect training AUCs, but the drops on the validation and test sets are large, which is indicative of the model overfitting. Naive Bayes performs well and consistently, while Decision Tree has the worst AUC performance due to its weak class-separation ability These results suggest that both LR and SVM are the best stand-alone predictive models actually. Meanwhile, boosting models have also gained strong prediction capability.

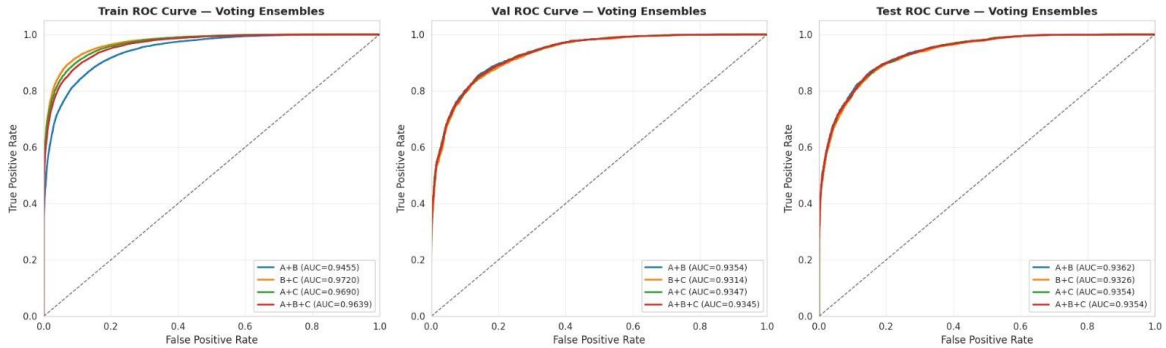

Fig. 8. ROC-AUC Curve (Ensemble Model)

Figure 8: ROC curves of voting-ensemble models on train, validation and test datasets (Author summary) All ensemble configurations are well above the random classification diagonal, confirming their efficiency in discriminating between flood-prone and non-flood-prone areas. Although B+C and A+C have better AUC scores in training, the performance gap becomes smaller in validation and test sets. A+B gives the best test AUC, but A+C and A+B+C produce nearly the same and similarly high results. The good agreement between validation and test performance suggests stable generalisation, and indicates that the voting ensembles are relatively insensitive to overfitting. In general, the SCA with multiple classifiers provides a stable and robust discrimination of flood risk for unseen samples.

D. Machine Learning Model Performance

TABLE II
MACHINE LEARNING MODEL PERFORMANCE

Model	Accuracy	Precision	Recall	F1-score
SVM	0.8517	0.8623	0.8443	0.8532
NB	0.8315	0.8455	0.8194	0.8323
KNN	0.7979	0.8337	0.7584	0.792
LR	**0.8573**	**0.8615**	**0.8584**	**0.8599**
DT	0.6735	0.6924	0.648	0.6695
RF	0.8273	0.8283	0.8346	0.8314
XGBoost	0.8457	0.853	0.843	0.8479
AdaBoost	0.8445	0.8502	0.844	0.8471
LightGBM	0.8435	0.8508	0.8406	0.8457
A+B	**0.8559**	**0.8636**	**0.8521**	**0.8578**
B+C	0.8511	0.8594	0.8466	0.853
A+C	0.8539	0.8617	0.85	0.8558
A+B+C	0.8555	0.8635	0.8513	0.8574

Table II compares nine base models and four voting ensembles for the flood-risk prediction. The best base model was Logistic Regression (LR) which gave an accuracy of 0.8573, precision of 0.8615, recall of 0.8584 and F1-score of 0.8599. The SVM was second with accuracy, precision, recall, and F1-score of 0.8517, 0.8623, 0.8443, and 0.8532, respectively. The XGBoost was third with accuracy, precision, recall, and F1-score of 0.8457, 0.8530, 0.8430, and 0.8479, respectively. AdaBoost and LightGBM also performed competitively with accuracies of 0.8445 and 0.8435 respectively. NB and RF have moderate performance and KNN has lower recall of 0.7584. The Decision Tree had the lowest performance, with an accuracy of 0.6735, precision of 0.6924, recall of 0.6480, and an F1-score of 0.6695.

A, B and C are LR, SVM and XGBoost respectively for ensemble construction. The highest ensemble performance was achieved by the A+B ensemble with 0.8559 accuracy, 0.8636 precision, 0.8521 recall and 0.8578 F1-score. A+B+C obtained almost similar results with 0.8555, 0.8635, 0.8513 and 0.8574 for accuracy, precision, recall and F1-score, respectively. The A+C achieved 0.8539, 0.8617, 0.8500 and 0.8558 values whereas the B+C achieved 0.8511 accuracy, 0.8594 precision, 0.8466 recall and 0.8530 F1-score. Overall, LR was the best model on its own and A+B was the best voting ensemble.

V. DISCUSSION

A+B was the best performing ensemble among the voting ensembles with 85.59% accuracy, 86.36% precision, 85.21% recall, 85.78% F1-score and the lowest ensemble MSE of 0.1441. Its confusion matrix had 3,261 true positives and 3,158 true negatives, with only 515 false positives, the joint-lowest FP among all models tested. The whole A+B+C combination also achieved competitive results with 85.55% accuracy and 85.74% F1-score. However the third model did not improve the performance over A+B . The extra classifier might have contributed to some redundant or less informative predictions . Although LR was 0.14 percentage points more accurate than A+B, A+B resulted in fewer false positives and offered a more balanced

trade-off between precision and recall. Therefore, A+B is the most reliable voting configuration, while LR is the best individual classifier.

These results are further validated by ROC analysis. LR achieved the highest test AUC of 0.9375 among the base models, followed by SVM (0.9321), XGBoost (0.9284), LightGBM (0.9277) and AdaBoost (0.9263). More importantly, LR yielded very consistent AUC values across training, validation and testing (0.9385, 0.9372 and 0.9375 respectively) indicating excellent generalization. For KNN and RF, the values of training AUC were 1.0000, but the test set were 0.8790 and 0.9141, respectively. This large train-test difference indicates overfitting, particularly for KNN. Decision Tree showed the worst ability of discrimination with a test AUC of 0.7334.

The voting ensembles demonstrated more stable ROC performance than some of the complex base classifiers. A+B obtained AUC values of 0.9455, 0.9354, and 0.9362 on training, validation, and test sets, respectively. It showed good generalization with very little overfitting (gap between train and test AUC was just 0.0093). A+B+C obtained test AUC of 0.9354 while A+C, B+C achieved 0.9354, 0.9326 respectively. Thus, while the ensembles did not significantly outperform LR in AUC, they consistently showed high discrimination on unseen validation and test data. The validation and test curves are also similar, which suggests that the reported results were not due to an unusually favourable test split.

The proposed results are compared favorably with several recent studies reported in the related work. Prakash et al. [14] used Random Forest for flood inundation mapping and reported a maximum overall accuracy of approximately 82.91%. In comparison, the proposed A+B ensemble achieved 85.59% accuracy, an improvement of about 2.68 percentage points while LR improved the result by about 2.82 percentage points. (Springer) Yosri et al. [24] reported an accuracy of about 80% for flood-risk-level classification using the RAPFLO framework. The proposed A+B and LR models outperformed this result by approximately 5.59 and 5.73 percentage points, respectively. But RAPFLO also dealt with future flood extent and probability, so the task was larger than the present binary classification problem. (Science Direct)

The conventional SVM baseline proposed by Srivastava et al. [6] only achieved an accuracy of 79.2% while the SVM implemented in this study achieved an accuracy of 85.17%, an improvement of about 5.97 percentage points. The proposed A+B ensemble also surpassed that SVM baseline by 6.39 percentage points. However, their optimized BO-EDNN achieved ultimately 95.2% accuracy, thus the present model outperformed their conventional baseline but not their final optimized deep-learning model. (Malque Similarly, Kafi et al. [22] reported an SVM ROC-AUC of 0.84, while the current SVM achieved a test AUC of 0.9321. This is equivalent to an increase in AUC of about 0.0921, but a direct comparison cannot be made as datasets, geographic conditions, predictor variables and validation methods are different. (ScienceDirect)

The results overall indicate that the proposed framework is a robust and balanced solution for flood classification. LR achieved the best numerical accuracy, recall, F1-score and test AUC.

A+B voting ensemble achieved the best ensemble accuracy, precision and lowest false positive count. The small differences between the training, validation and test AUC values show that LR and A+B generalized more consistent than KNN, RF, XGBoost and LightGBM. Therefore, A+B is suitable when the decrease of false alarms and the constancy of the performance along the cross-split are more important, whereas LR is suitable when the maximization of the overall classification accuracy and the identification of the largest number of flood cases are the main objectives.

However, the comparison with previous studies should be taken with caution as the studies used different datasets, target definitions, geographical areas, class distributions and evaluation metrics. Accuracy, AUC, R-squared, NSE and hit rate are not directly interchangeable. Moreover, the reported MSE values are precisely 1 minus accuracy, which is a clear indication that MSE was computed on the basis of hard class labels. Future evaluation should include probability-based measures such as the Brier score, log loss, calibration curves and external geographical validation to establish the practical reliability of the proposed model for flood early-warning applications in the real world.

VI. CONCLUSION

In this study, nine base machine-learning classifiers and four voting ensemble configurations were evaluated for flood classification using the accuracy, precision, recall, F1-score, MSE, confusion-matrix outcomes, and ROC-AUC. Among the individual models, Logistic Regression performed the best overall with an accuracy of 85.73%, recall of 85.84%, F1-score of 85.99% and the highest test AUC of 0.9375. Its close training, validation, and test AUC values also indicated good generalization and little sign of overfitting. However, KNN and Random Forest have obvious train-test gaps, and Decision Tree has the worst classification and discrimination results.

The most efficient configuration among the voting ensembles was A+B, which achieved 85.59% accuracy, 86.36% precision, 85.21% recall, 85.78% F1-score, test AUC of 0.9362 and the lowest ensemble MSE of 0.1441. This also led to a small number of false positives (515) indicating a good trade-off between flood detection and false-alarm reduction. The third classifier in A+B+C did not improve performance. Thus, increasing the complexity of the ensemble does not necessarily improve the predictions, if the added model provides redundant or less accurate information.

Comparison with some recent studies on flood prediction reporting accuracy or AUC values less than 85% also proves that the proposed models perform well. However, such comparisons should be taken with caution, as different datasets, target definitions, geographical settings and evaluation procedures were used in previous studies. Overall, the Logistic Regression is the more appropriate model when the highest accuracy, recall, F1-score and AUC are desired, whereas the A+B ensemble is better if consistent results and fewer false positives are required. Additional work should include external geographical validation, real-time hydrological data, probability calibration, Brier score, log loss, and broader cross-regional testing to demonstrate the practical applicability of the framework in operational flood early-warning systems.

REFERENCES

[1] World Health Organization, *"Climate Change and Health,"* Geneva, Switzerland, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/climate-change-and-health

[2] United Nations Office for Disaster Risk Reduction, *Global Assessment Report on Disaster Risk Reduction 2023: Our World at Risk*, Geneva, Switzerland, 2023. [Online]. Available: https://www.undrr.org/gar

[3] Intergovernmental Panel on Climate Change, *Climate Change 2022: Impacts, Adaptation and Vulnerability*. Cambridge, U.K.: Cambridge University Press, 2022, doi: 10.1017/9781009325844. [Online]. Available: https://doi.org/10.1017/9781009325844

[4] World Bank, *"Climate Change and Disaster Risk Management,"* Washington, DC, USA, 2022. [Online]. Available: https://www.worldbank.org/en/topic/disasterriskmanagement

[5] United Nations Environment Programme, *Adaptation Gap Report 2023*. Nairobi, Kenya, 2023. [Online]. Available: https://www.unep.org/resources/adaptation-gap-report-2023

[6] R. Srivastava, S. Pathak, Lakshmanan, D. Sharma, B. Bhushan, and Mahalakshmi, "Novel data-driven model for flood prediction using big data: Social resilience and risk management," *Multidisciplinary Science Journal*, vol. 7, Art. no. e2025ss0315, 2025, doi: 10.31893/multiscience.2025ss0315.

[7] L. Cui, G. Hu, and Y. Zhu, "Multi-strategy improved snow ablation optimizer: A case study of optimization of kernel extreme learning machine for flood prediction," *Artificial Intelligence Review*, vol. 58, Art. no. 181, 2025, doi: 10.1007/s10462-025-11192-z.

[8] H. Mancy, N. E. Ghannam, A. Abozeid, and A. I. Taloba, "Decentralized multi-agent federated and reinforcement learning for smart water management and disaster response," *Alexandria Engineering Journal*, vol. 126, pp. 8–29, 2025, doi: 10.1016/j.aej.2025.04.033.

[9] M. Umadevi *et al*., "Design of mTCN framework for disaster prediction: A fusion of massive machine type communications and temporal convolutional networks," *Scientific Reports*, vol. 15, Art. no. 29280, 2025, doi: 10.1038/s41598-025-15397-8.

[10] Y. Irawan, R. Wahyuni, and Herianto, "An integrated machine learning and deep learning approach for multiclass flood risk classification with feature selection and imbalanced data handling," *Jurnal Teknologi dan Open Source*, vol. 8, no. 2, pp. 650–661, 2025, doi: 10.36378/jtos.v8i2.4639.

[11] R. Sovia *et al*., "MLSEL: A tuned multilayer stacking ensemble learning with meta feature for flood risk prediction," *Engineering and Applied Science Research*, vol. 53, no. 2, pp. 211–222, 2026, doi: 10.64960/easr.2026.263071.

[12] Y. Wang, J. Zhang, and Y. Chang, “A probability prediction model for flood disasters based on multi-layer perceptron,” *Journal of Physics: Conference Series*, vol. 2905, Art. no. 012003, 2024, doi: 10.1088/1742-6596/2905/1/012003.

[13] Y. Zhang, Y. Kang, Y. Liu, R. Fan, and J. Duan, “Study on flood probability by neural network prediction based on the importance of random forest characteristics,” in *Proc. SPIE*, vol. 13512, Art. no. 135120R, 2025, doi: 10.1117/12.3057166.

[14] A. J. Prakash *et al*., “Development of an automated method for flood inundation monitoring, flood hazard, and soil erosion susceptibility assessment using machine learning and AHP–MCE techniques,” *Geoenvironmental Disasters*, 2024, doi: 10.1186/s40677-024-00275-8.

[15] M. E. S. El-Mahdy *et al*., “Flood classification and prediction in South Sudan using artificial intelligence models under a changing climate,” *Alexandria Engineering Journal*, 2024, doi: 10.1016/j.aej.2024.03.082.

[16] Y. Hahn *et al*., “Towards accurate flood predictions: A deep learning approach using Wupper River data,” *Water*, vol. 16, no. 23, Art. no. 3368, 2024, doi: 10.3390/w16233368.

[17] M. G. Alam, V. Tripathi, C. M. Bhatt, and M. P. Mohanty, “A novel framework embedding Bayesian-optimized ensemble machine learning and explainable artificial intelligence to improve flood prediction in complex watersheds,” *Environmental and Sustainability Indicators*, vol. 27, Art. no. 100760, 2025, doi: 10.1016/j.indic.2025.100760.

[18] C. Singha *et al*., “A novel framework for flood susceptibility assessment using hybrid analytic hierarchy process-based machine learning methods,” *Natural Hazards*, vol. 121, pp. 13765–13810, 2025, doi: 10.1007/s11069-025-07335-8.

[19] S. Xu *et al*., “Urban flood prediction model based on explainable deep learning and attention mechanism,” *International Journal of Disaster Risk Science*, vol. 17, pp. 161–181, 2026, doi: 10.1007/s13753-026-00691-4.

[20] M. I. Hossain *et al*., “Geospatial driven machine learning approach for flood susceptibility mapping in southeastern region of Bangladesh,” *Discover Geoscience*, vol. 4, Art. no. 201, 2026, doi: 10.1007/s44288-026-00572-9.

[21] S. Chardi *et al*., “Machine learning-based flood susceptibility mapping in the Ourika Watershed, Morocco,” *Journal of Mountain Science*, vol. 23, pp. 2472–2489, 2026, doi: 10.1007/s11629-025-0393-7.

[22] K. M. Kafi, Z. Ponrahono, Z. H. Ash’aari, and A. S. Barau, “Flood risk prediction and modeling in Bauchi: Leveraging machine learning models and explainable AI for urban resilience,” *Journal of Climate Change and Health*, vol. 26, Art. no. 100490, 2025, doi: 10.1016/j.joclim.2025.100490.

[23] M. Wahba *et al*., “Examination of the efficacy of machine learning approaches in the generation of flood susceptibility maps,” *Environmental Earth Sciences*, vol. 83, Art. no. 429, 2024, doi: 10.1007/s12665-024-11696-x.

[24] A. Yosri, M. Ghaith, and W. El-Dakhakhni, “Deep learning rapid flood risk predictions for climate resilience planning,” *Journal of Hydrology*, vol. 631, Art. no. 130817, 2024, doi: 10.1016/j.jhydrol.2024.130817.

25] M. E. Chowdhury, A. K. M. S. Islam, R. U. Zzaman, and S. Khadem, “A machine learning-based approach for flash flood susceptibility mapping considering rainfall extremes in the northeast region of Bangladesh,” *Advances in Space Research*, vol. 75, no. 2, pp. 1990–2017, 2025, doi: 10.1016/j.asr.2024.10.047.

[26] G. Zhao *et al*., “Large-scale flash flood warning in China using deep learning,” *Journal of Hydrology*, vol. 604, Art. no. 127222, 2022, doi: 10.1016/j.jhydrol.2021.127222.

[27] C. Chen, J. Jiang, Z. Liao, Y. Zhou, H. Wang, and Q. Pei, “A short-term flood prediction based on spatial deep learning network: A case study for Xi County, China,” *Journal of Hydrology*, vol. 607, Art. no. 127535, 2022, doi: 10.1016/j.jhydrol.2022.127535.

[28] Z. Wu, Y. Zhou, H. Wang, and Z. Jiang, “Depth prediction of urban flood under different rainfall return periods based on deep learning and data warehouse,” *Science of the Total Environment*, vol. 716, Art. no. 137077, 2020, doi: 10.1016/j.scitotenv.2020.137077.

[29] R. Kondo, B. Du, Y. Narusue, and H. Morikawa, “Machine learning framework supervised by hydraulic mechanical models for real-time pluvial flood prediction,” *Journal of Information Processing*, vol. 31, pp. 256–264, 2023, doi: 10.2197/ipsjjip.31.256.

[30] A. Mosavi, P. Ozturk, and K.-W. Chau, “Flood prediction using machine learning models: Literature review,” *Water*, vol. 10, no. 11, Art. no. 1536, 2018, doi: 10.3390/w10111536.