Predictive Flood Susceptibility Modelling with Machine Learning: Insights from The Baitarani River Basin, Odisha

Dillip Kumar Ghose[1*], *Bibhu Prasad Mishra*[2], and *Sourav Ghose*[3]

[1]Department of Civil Engineering, National Institute of Technology, Silchar, Assam, India, 788010
[2]School of Infrastructure and Planning, Odisha University of Technology and Research, Bhubaneswar, Odisha, India, 751029
[3]Graduate Engineer Trainee (Civil), Utkal Alumina International Ltd., Odisha, India, 765015

Abstract. Floods are among the most destructive natural hazards, causing widespread damage to ecosystems, communities, and critical infrastructure. Accurately understanding the drivers of flooding and identifying areas prone to inundation are essential for informed disaster management and planning. This study analyzes historical flood records from 2003 to 2023 to delineate flood-susceptible (FS) zones within the Baitarani River Basin (BRB). To predict these vulnerable areas, two machine learning (ML) models were applied: the Fuzzy Support Vector Machine (FSVM) and an enhanced version optimized using Simulated Annealing (SA-FSVM). These models were developed to capture the complex relationships between flood events and multiple contributing environmental factors. Model performance was evaluated using the Area Under the Receiver Operating Characteristic (AUROC) curve, a robust metric for assessing classification accuracy and reliability. Among the tested models, SA-FSVM achieved the highest predictive performance, with an AUROC value of 0.91. The findings reveal that low-lying coastal zones, particularly in the southeastern portions of the basin, are highly vulnerable to flooding. Overall, the study demonstrates the capability of advanced ML techniques for effective flood susceptibility mapping and highlights their importance in guiding flood mitigation strategies and enhancing disaster preparedness across the BRB.

Keywords: Flood susceptibility, Baitarani River Basin, Machine Learning, GIS, Receiver Operating Characteristic

1 INTRODUCTION

One common and destructive natural disaster is flooding which is particularly common in river basins like the Baitarani River Basin (BRB) in Odisha India. Because of its location topography and climate, the BRB frequently experiences severe flooding. Implementing

* Corresponding author: dillip@civil.nits.ac.in

efficient flood management and mitigation strategies in this region depends on understanding and mapping flood susceptibility. Beginning in the Guptaganga Hills in the Keonjhar district of Odisha the Baitarani River eventually flows into the Bay of Bengal. The socioeconomic fabric of the area is greatly impacted by the frequent and severe flooding caused by its distinctive drainage features. Because of the location of the river and its reliance on precipitation flooding in the BRB is frequent and erratic which is why local communities are very concerned about it. Floods are more frequent and severe when severe monsoonal rainfall is combined with drainage congestion. One of the main causes of these floods is the basins prolonged and intense rainfall. Historically, the area has seen devastating floods in years like 2006 and 2011, which have strengthened the river's reputation for being unpredictable and causing extensive damage.

Tree classifiers, hybrid support vector machines deep learning techniques Naïve Bayes Tree (NBT) Naïve Bayes (NB) and other machine learning algorithms have all been used for flood-susceptible (FS) mapping worldwide [1–6]. By combining various data sources, these methods increase the accuracy of flood predictions. Though FS mapping has advanced traditional methods frequently overlook complex variables and their interactions leading to forecasts that are inconsistent. By creating ML models that can precisely identify areas that are prone to flooding this study overcomes these constraints. To improve FS mapping the study looks into topography land use hydrological conditions and past flood events. This approach seeks to promote proactive flood risk management and better-informed decision-making. ML is especially useful for this purpose because of its capacity to reveal complex patterns in big datasets as shown in research by [7–11].

This study develops ML models, that is to say, Fuzzy Support Vector Machine (FSVM) and Simulated Annealing-optimized FSVM (SA-FSVM), for the identification of flood-prone zones in BRB basing on historical data of floods. To what extent these models can recognize the patterns of floods, enhance the accuracy of prediction, and improve FS maps with data from regions that are flood-prone is what this study endeavors to assess. The approach thus proposed shall try to overcome the limitations conventional methods of FS mapping have and bring to bear more reliable tools in flood risk administration.

2 METHODOLOGY

2.1 Study Area

Situated in Odisha India BRB occupies an area of roughly 13482 square kilometers. Its latitude is between 20°28N and 23°38N and its longitude is between 83°55E and 87°03E. The basins' varied topography which includes both flat alluvial plains and hilly terrains has a big impact on the rivers flood dynamics and hydrological behavior. With significant rainfall from June to September the BRB has a tropical monsoon climate. Because the area receives between 1500 and 2200 mm of rainfall annually the basin is extremely vulnerable to flooding. Water bodies cities forests and agriculture make up the BRBs land use. While forests mostly cover the upper reaches agriculture especially rice cultivation predominates in the lower plains. But in downstream regions growing urbanization and industrial activity increase the risk of flooding by lowering soil permeability and increasing surface runoff. The river is vital to the local population's industrial agricultural and fishing operations but frequent floods pose serious problems for ecosystems infrastructure and livelihoods. Through an analysis of topographical hydrological and land use factors this study seeks to improve flood risk management in the BRB and increase community resilience. The foundation of the spatial analysis is the Baitarani sub-basin shapefile which was acquired from India's Open Government Data (OGD) platform (Figure 1).

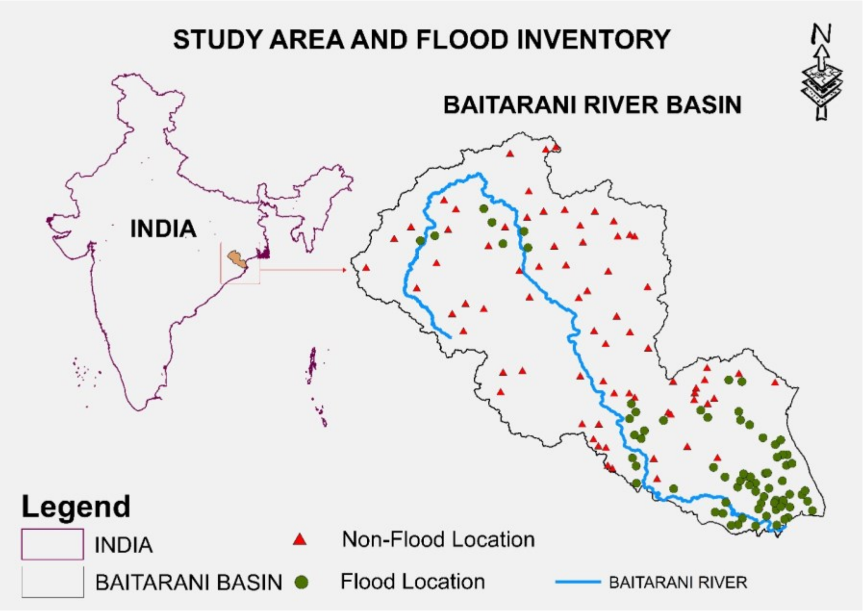

Fig. 1. Study area and flood inventory map.

2.2 Flood Inventory map

As stated by [12], the development of a flood susceptibility map requires the creation of a flood inventory map. The Bhuvan portal of the Indian Space Research Organization (ISRO) provided the flood inventory data used in this study which included details on 146 areas affected by flooding. Seventy percent (102 points) of these sites were used for training and thirty percent (44 points) were used for testing FS models (Fig. 1. Whereas non-flood locations indicate areas where there have been no recorded flood events in the previous 20 years flood locations represent specific areas with multiple known flooding events. Using a binary encoding system flood locations were given a value of 1 and non-flood locations a value of 0. According to the guidelines of [13] the number of non-flood points was maintained at the same level as the number of flood points in order to reduce sampling bias. Using historical flood records spatial analysis and expert knowledge non-flood locations were determined to ensure that they accurately represented areas that were not impacted by flooding during the study period. Machine learning models were trained and their efficacy assessed using the dataset which contained an equal number of points representing flood and non-flood areas. The model's robustness and dependability are improved by this balanced dataset approach which lowers the possibility of overfitting or underestimating susceptibility predictions. In order to create spatially explicit FS maps and integrate flood conditioning factors the flood inventory map is a fundamental input. These maps are essential for locating high-risk areas and assisting with well-informed flood management and mitigation decision-making.

2.3 Flood Conditioning factors

A number of important conditioning factors were examined in order to evaluate FS in the Baitarani River Basin (BRB). Information about elevation above sea level which has a direct impact on water flow and the likelihood of flooding is provided by elevation data. The rate and amount of surface runoff are influenced by the slope or steepness of the land. A slopes aspect or the direction it faces affects the distribution of sunlight evaporation rates soil moisture and vegetation growth. Convex and concave features change the direction and speed of water flow across the terrain and curvature describes the shape of the land surface. Rainfall data are crucial for determining how much precipitation causes flooding and rainfall intensity tells us how quickly precipitation builds up over time which has a big influence on runoff rates and flood severity. Using information from the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS) and the Google Earth Engine platform rainfall intensity maps were produced. The hydrological cycle and flood risks are significantly impacted by temperature which also affects evaporation snowmelt and soil moisture retention. The density and health of the vegetation which affect runoff and water absorption were evaluated using the Normalized Difference Vegetation Index (NDVI). By increasing soil permeability densely vegetated areas can reduce the risk of flooding. By measuring the erosive force of flowing water, the Stream Power Index (SPI) sheds light on how floods alter the terrain. Similar to this the Sediment Transport Index (STI) evaluates the movement of sediment during floods and provides important data on soil erosion and displacement. To better understand soil moisture and possible flood risks the Topographic Wetness Index (TWI) highlights regions that are susceptible to water saturation. The Terrain Ruggedness Index (TRI) gauges the variability and roughness of the landscape which have an impact on water accumulation and flow. Data on Land Use and Land Cover (LULC) shed light on the ways that both natural and man-made environments affect the dynamics of floods. Urban areas farms and forests are examples of land use types that have varying effects on water runoff and infiltration (Fig. 2). For example while forests and vegetated areas encourage absorption and reduce the risk of flooding urbanization increases impervious surfaces which raises runoff rates. Advanced machine learning models were used to combine these factors in order to more accurately predict areas that are prone to flooding. Through the investigation of the complex interrelationships among flood conditioning variables these models offer more profound understanding of flood risk trends throughout the BRB. By helping to develop more effective flood management plans this method lessens the detrimental effects of flooding on nearby communities and vital infrastructure.

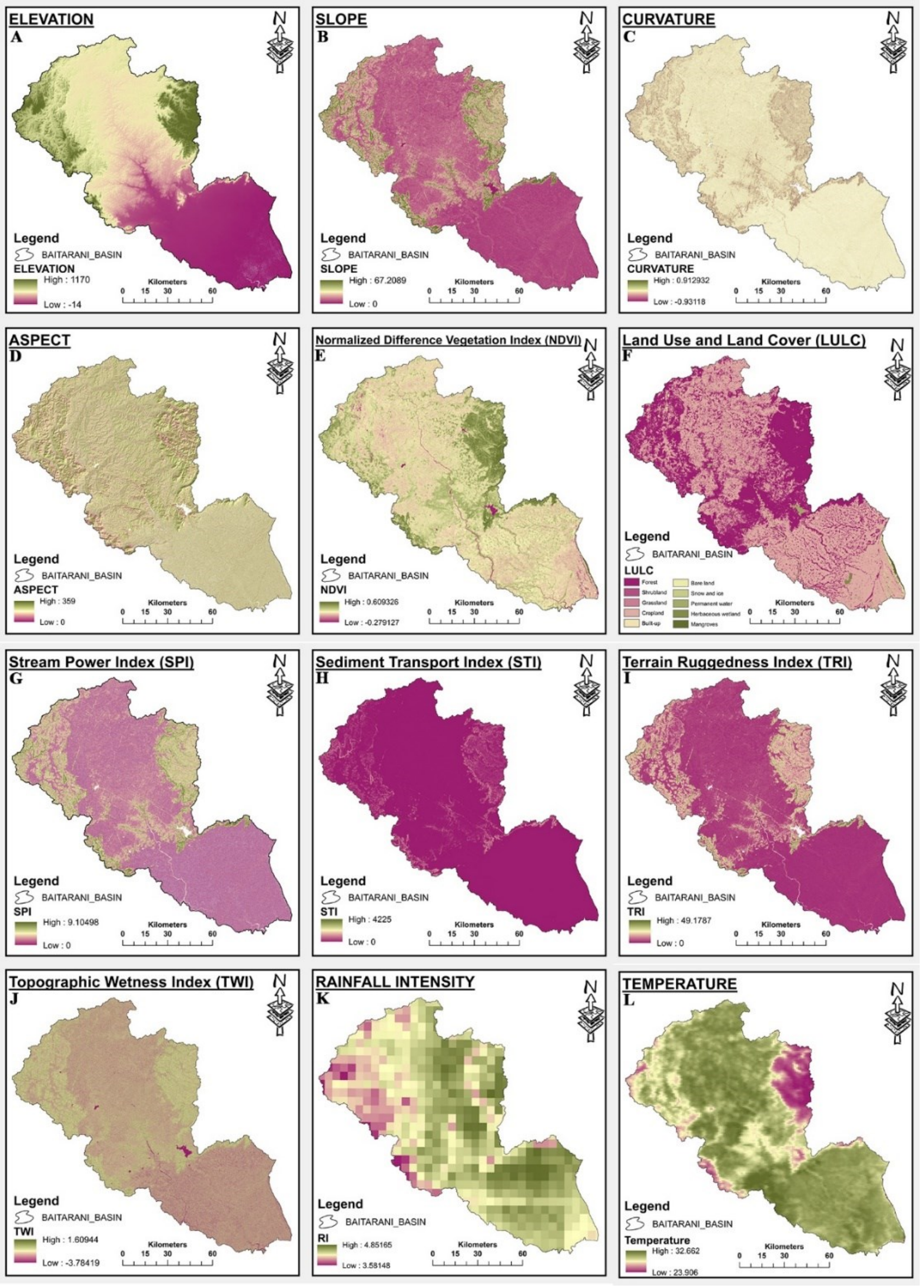

Fig. 2. Flood conditioning factors.

2.4 FSVM for FS Mapping

The FSVM assigns fuzzy membership values to each sample, reducing the impact of noisy or ambiguous data [14]. For a given sample (x_i, y_i), where x_i is feature and $y_i \in \{-1, +1\}$ is class label, fuzzy membership μ_i is calculated as:

$$\mu_i = 1 - \frac{d\left(x_i, C_{flood}\right)}{d\left(x_i, C_{flood}\right) + d\left(x_i, C_{non-flood}\right)}$$

where $d\left(x_i, C_{flood}\right)$ and $d\left(x_i, C_{non-flood}\right)$ are the distances of x_i from the flood-prone and non-flood-prone class centers, respectively.Objective of FSVM is to minimize the following loss function:

subject to:

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1,2, \dots, N$$

Here, w is weight vector, bb is bias term, ξ_i are slack variables, CC is regularization parameter, and $\phi(x_i)$ is the kernel function.

2.5 Simulated Annealing for Optimized FSVM

Simulated Annealing (SA) [15] was used to optimize the hyperparameters of the FSVM, including C, γ (kernel parameter), and fuzzy membership parameters. The SA algorithm iteratively generates new parameter sets, evaluating them using: $P = \exp\left(-\frac{\Delta E}{T}\right)$ where ΔE is the difference in performance between current and new solutions, and T is temperature [16]. The temperature T is reduced at each iteration using: $T_{k+1} = \alpha T_k, \quad \alpha \in (0,1)$. The optimized parameters were then used to train the FSVM model.

2.6 Model Validation

The FSVM and SA-FSVM models' performance was evaluated using the AUROC curve [17]. The calculation of AUROC is presented as follows:

$$AUROC = \int_0^1 TPR(FPR)\ d(FPR)$$

Where TPR (True Positive Rate) and FPR (False Positive Rate) are derived from the confusion matrix.

The models were also evaluated using accuracy, sensitivity, specificity, and other statistical metrics.

3 RESULTS

3.1 Analysis of Flood Susceptible Map

An essential tool for lowering flood vulnerability and putting preventative measures into action is the FS map which is produced during the modeling process. The most vulnerable areas are graphically highlighted on this map allowing for prompt identification and focused interventions. Based on the probability of flooding there are four levels of flood risk: very low moderate high and very high. ArcGIS 10. 8s manual break method is used to apply these classifications to the FS map (Fig. 3. Coastal and low-lying southeast regions which are known for their low elevation and historical vulnerability to flooding were consistently found to be extremely flood-prone on all maps produced by the model.

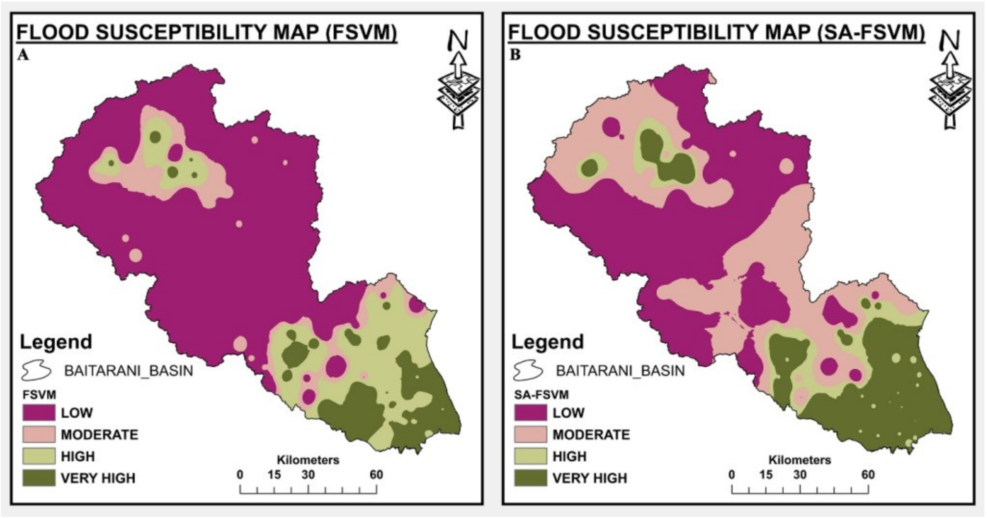

Fig. 3. Flood Susceptibility Map.

3.2 Model Performance Comparison

In machine learning (ML) hyperparameter tuning is the process of choosing the best combination of parameters to improve model performance. To evaluate different parameters and determine the ideal values for flood susceptibility models this study employed a theoretical framework. The dataset was divided with 30% set aside for testing and 70% designated for training. Using the AUROC metric which gauges the model's efficacy and forecasting ability the predictive accuracy of models in identifying flood-prone areas was evaluated. Excellent performance is indicated by AUC values between 0 and 1 where higher values signify greater accuracy. By increasing the precision and dependability of FS mapping this strategy sought to facilitate better flood risk management. The AUROC metric was used to compare the performance of the Simulated Annealing-optimized FSVM (SA-FSVM) and Fuzzy Support Vector Machine (FSVM) models (Fig. 4). Superior performance was shown by the SA-FSVM which outperformed the FSVM by about 4% in terms of AUROC.

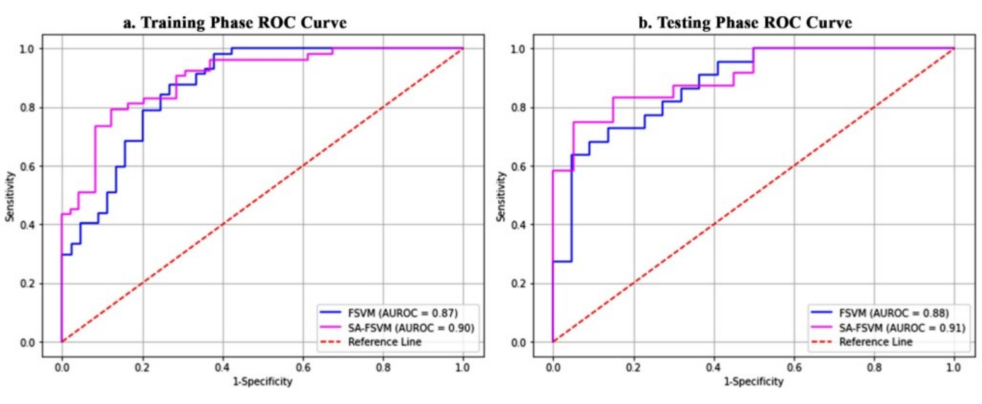

Fig. 4. AUROC curve of a. Training and b. Testing phase.

4 DISCUSSION

Severe flooding occurs regularly in the BRB, underscoring the need for efficient planning and management techniques. Developing a flood susceptibility (FS) map is essential for reducing these impacts. To capture the spatial distribution and frequency of past flood events, flood frequency maps were generated using data from the past two decades. The analysis shows that medium-frequency floods affect roughly 8–68% of the basin, while high-frequency floods impact about 6–46% of the area. These maps play a crucial role in flood control and proactive planning, helping authorities minimize future damage. The modelling approaches used in this study were practical, interpretable, and suitable for large-scale FS assessment. However, the lack of detailed hydrological datasets, such as flood depth, flow velocity, and discharge, limited the development of a more comprehensive predictive model. Despite the inherent complexity and uncertainty associated with flood modelling, machine learning methods have demonstrated strong potential to overcome these challenges, especially when reliable historical flood inventory data are available [18,19].

5 CONCLUSION

The research methodology used in this study has several benefits because it estimates flood-prone areas without the need for intricate hydrological models. The practical application of traditional hydrological models is limited because they frequently call for enormous volumes of detailed data, which can be challenging to acquire. By contrast, machine learning techniques effectively pinpoint areas at risk of flooding by utilizing easily accessible data including elevation LULC and past flood occurrences. The process is made simpler by this method which also expands the range of applications beyond what is usually possible with traditional hydrological models. ML algorithms are especially useful for mapping flood susceptibility because of their exceptional ability to identify intricate patterns and connections in large datasets. Even with a few data inputs these models can process a wide range of complex variables and produce precise predictions. In areas with limited or non-existent detailed hydrological data, this capability is especially helpful. Furthermore, the adaptability of machine learning techniques makes them applicable in a variety of climatic and geographic contexts increasing their usefulness for various research fields. Future research should focus on integrating ML techniques with hydrological models to enhance flood prediction accuracy. Combining these methods can offer a more comprehensive understanding of flood vulnerability, especially when key hydrological parameters such as flood depth, velocity, and discharge are incorporated. Hybrid approaches—for instance, using artificial neural networks alongside hydrological simulations—can help address limitations observed in standalone models. Further studies should also examine how land use changes, including urban expansion and deforestation, influence flood susceptibility. These shifts significantly affect runoff patterns and flood behavior, making their inclusion in ML-based assessments essential. Incorporating land use dynamics can provide planners and policymakers with valuable insights for developing effective flood mitigation strategies and strengthening community resilience. Overall, integrating advanced ML methods with traditional hydrological modeling, while accounting for land use changes, can substantially improve flood risk assessments and support more efficient and proactive flood management practices.

References

1. K. Khosravi, M. Panahi, A. Golkarian, S.D. Keesstra, P.M. Saco, D.T. Bui, S. Lee, Convolutional neural network approach for spatial prediction of flood hazard at national scale of Iran, J Hydrol (Amst) 591 (2020) 125552. https://doi.org/10.1016/j.jhydrol.2020.125552.
2. S. Talukdar, B. Ghose, Shahfahad, R. Salam, S. Mahato, Q.B. Pham, N.T.T. Linh, R. Costache, M. Avand, Flood susceptibility modeling in Teesta River basin, Bangladesh using novel ensembles of bagging algorithms, Stochastic Environmental Research and Risk Assessment 34 (2020) 2277–2300. https://doi.org/10.1007/s00477-020-01862-5.
3. B.P. Mishra, D.K. Ghose, D.P. Satapathy, Geospatial modeling using hybrid machine learning approach for flood susceptibility, Earth Sci Inform 15 (2022) 2619–2636. https://doi.org/10.1007/s12145-022-00872-x.
4. D.P. Satapathy, B.P. Mishra, Flood susceptibility modeling by integrating tree-based regression with metaheuristic algorithm: BWO, Transactions in GIS (2024). https://doi.org/10.1111/tgis.13171.
5. Y. Wang, Z. Fang, H. Hong, L. Peng, Flood susceptibility mapping using convolutional neural network frameworks, J Hydrol (Amst) 582 (2020) 124482. https://doi.org/10.1016/j.jhydrol.2019.124482.
6. V. Shivhare, A. Kumar, R. Kumar, S. Shashtri, J. Mallick, C.K. Singh, Flood susceptibility and flood frequency modeling for lower Kosi Basin, India using AHP and Sentinel-1 SAR data in geospatial environment, Natural Hazards (2024) 1–32. https://doi.org/10.1007/s11069-024-06614-0.
7. Z. Wang, C. Lai, X. Chen, B. Yang, S. Zhao, X. Bai, Flood hazard risk assessment model based on random forest, J Hydrol (Amst) 527 (2015) 1130–1141. https://doi.org/10.1016/j.jhydrol.2015.06.008.
8. R. Costache, L. Zaharia, Flash-flood potential assessment and mapping by integrating the weights-of-evidence and frequency ratio statistical methods in GIS environment – case study: Bâsca Chiojdului River catchment (Romania), Journal of Earth System Science 126 (2017) 59. https://doi.org/10.1007/s12040-017-0828-9.
9. R. Madhuri, S. Sistla, K. Srinivasa Raju, Application of machine learning algorithms for flood susceptibility assessment and risk management, Journal of Water and Climate Change 12 (2021) 2608–2623. https://doi.org/10.2166/wcc.2021.051.
10. S.V. Razavi-Termeh, A. Pourzangbar, A. Sadeghi-Niaraki, M.J. Franca, S.-M. Choi, Metaheuristic-driven enhancement of categorical boosting algorithm for flood-prone areas mapping, International Journal of Applied Earth Observation and Geoinformation 136 (2025) 104357. https://doi.org/10.1016/j.jag.2025.104357.
11. B.A. El-Haddad, A.M. Youssef, A.M. Mahdi, Z. Karimi, H.R. Pourghasemi, National flood susceptibility mapping in Saudi Arabia, Earth Sci Inform 18 (2025) 118. https://doi.org/10.1007/s12145-024-01510-4.
12. S. Saravanan, D. Abijith, Flood susceptibility mapping of Northeast coastal districts of Tamil Nadu India using Multi-source Geospatial data and Machine Learning techniques, Geocarto Int 37 (2022) 15252–15281. https://doi.org/10.1080/10106049.2022.2096702.
13. A.R.M. Towfiqul Islam, S. Talukdar, S. Mahato, S. Kundu, K.U. Eibek, Q.B. Pham, A. Kuriqi, N.T.T. Linh, Flood susceptibility modelling using advanced ensemble

machine learning models, Geoscience Frontiers 12 (2021) 101075. https://doi.org/10.1016/j.gsf.2020.09.006.
14. Chun-Fu Lin, Sheng-De Wang, Fuzzy support vector machines, IEEE Trans Neural Netw 13 (2002) 464–471. https://doi.org/10.1109/72.991432.
15. D. Bertsimas, J. Tsitsiklis, Simulated Annealing, Statistical Science 8 (1993). https://doi.org/10.1214/ss/1177011077.
16. R.A. Rutenbar, Simulated annealing algorithms: an overview, IEEE Circuits and Devices Magazine 5 (1989) 19–26. https://doi.org/10.1109/101.17235.
17. C. Marzban, The ROC Curve and the Area under It as Performance Measures, Weather Forecast 19 (2004) 1106–1114. https://doi.org/10.1175/825.1.
18. S. Janizadeh, M. Avand, A. Jaafari, T. Van Phong, M. Bayat, E. Ahmadisharaf, I. Prakash, B.T. Pham, S. Lee, Prediction success of machine learning methods for flash flood susceptibility mapping in the Tafresh watershed, Iran, Sustainability 11 (2019) 5426. https://doi.org/https://doi.org/10.3390/su11195426.
19. T. Islam, E.B. Zeleke, M. Afroz, A.M. Melesse, A Systematic Review of Urban Flood Susceptibility Mapping: Remote Sensing, Machine Learning, and Other Modeling Approaches, Remote Sens (Basel) 17 (2025) 524. https://doi.org/10.3390/rs17030524.