

Predicting Riverbank Erosion Susceptibility Using Machine Learning and GIS

maryam ilanloo

`m.ilanloo6101@iau.ir`

Islamic Azad University

Research Article

Keywords: Riverbank erosion, machine learning, Random Forest, Geographic Information System (GIS), Sefidrood watershed

Posted Date: October 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11005859/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

The present study, with the aim of preparing a riverbank erosion sensitivity map in the Sefidrood watershed, used a combination of geographic information systems (GIS) and machine learning algorithms including Random Forest, XGBoost, LightGBM, and CatBoost. For this purpose, 14 factors affecting erosion including topographic, hydrological, geological, climatic, and land cover variables were extracted and after preparing the database, the models were trained using real data. The performance of the models was evaluated using the Accuracy, Precision, Recall, F1-Score, and ROC-AUC indices. The results showed that the Random Forest model with accuracy of 93.48%, precision of 91.50%, sensitivity of 88.66%, provided the best performance among the models studied. Also, the analysis of the importance of variables using the SHAP method showed that variables such as precipitation, flow accumulation, stream power index (SPI), lithology and vegetation cover index (NDVI) play the most role in controlling the sensitivity of riverbank erosion. The final sensitivity map indicates that the areas with high and very high sensitivity are mainly concentrated along the main river course and especially in the downstream and northern parts of the basin, which is consistent with the geomorphological and hydrological conditions of the region. The results of this study show that combining the optimized Random Forest algorithm with GIS is an efficient and reliable tool for predicting and zoning the susceptibility of riverbank erosion.

1.Introduction

Rivers are among the most dynamic geomorphological systems on the Earth's surface, constantly influenced by climate change, interactions of hydrological processes, sediment transport, tectonic activities, and human interventions. The interaction of these factors controls the stability of riverbanks, river channel pattern and shape, the development of alluvial plains, and sedimentation and erosion processes, playing a crucial role in agricultural land productivity, river ecosystem sustainability, socio-economic development of surrounding areas, and water resources management. In recent decades, due to the intensification of climate change, dam construction, expansion of agricultural activities, rapid urbanization, indiscriminate riverbed material extraction, and land use changes, the intensity of river morphological changes has increased in many parts of the world (Acquaotta et al., 2019; Begam & Sen, 2019; Jonathan & Charles, 2025).

One of the most important geomorphological hazards is riverbank erosion, which, in addition to reducing fertile lands and agricultural areas, causes the destruction of civil infrastructure, industrial facilities, and human settlements. On the other hand, sedimentation also accompanies changes in channel geometry, increased flood risk and reduced flow conveyance capacity, and changes in the ecological function of the river. Therefore, the simultaneous investigation of sedimentation and erosion processes as one of the most important indicators of river system dynamics has particular importance in sustainable watershed management planning (Rahman et al., 2024; Nguyen et al., 2026).

In recent years, climate change has intensified the morphological instability of rivers through increased occurrence of severe floods, long-term droughts, frequency of intense and torrential rainfall, and rising sea levels. These changes have not only altered sediment transport patterns but have also caused bank retreat, river channel migration, destruction or development of floodplains, and changes in coastal morphology. In rivers flowing into the sea, the entry of saline water, changes in tidal regimes, and reduced sediment supply due to dam construction have added to the complexity of these processes (Acquaotta et al., 2019; Tzeppenlis et al., 2022).

Alongside natural factors, human activities have also become one of the most important controlling factors of river morphological changes. Urban development, expansion of agricultural lands, dam construction, sand and gravel mining, dredging, construction of hydraulic structures, and land use changes have disrupted the natural balance of sediment transport and intensified erosion or sedimentation in different parts of rivers (Ebrahimi et al., 2025). Numerous studies in different parts of the world have shown that these human interventions play a decisive role in river morphological changes (Jonathan & Charles, 2025; Ilanloo, 2025).

Accurate investigation of these changes requires methods that can monitor the temporal and spatial evolution of rivers with high precision. Although traditional field survey methods have appropriate accuracy, they are not very suitable for long-term and extensive studies due to being time-consuming, high cost, and limitations in spatial coverage. Advances in Geographic Information Systems (GIS), remote sensing technology, and satellite imagery have enabled continuous monitoring of river morphological changes at different spatial and temporal scales (Tzeppenlis et al., 2022; Rahman et al., 2024; Soleimani & Ilanloo, 2025).

Multi-temporal satellite images, especially Landsat and Sentinel data, have today become one of the most important information sources for identifying riverbank changes, channel migration, floodplain changes, sedimentation, and land use changes. The integration of these data with GIS capabilities has enabled the extraction of quantitative indicators of morphological changes and analysis of river evolution trends with appropriate accuracy (Begam & Sen, 2019; Ebrahimi et al., 2025).

Numerous studies in recent years have used these technologies to investigate river morphological changes. For example, Jonathan and Charles (2025) used geospatial data and machine learning algorithms to evaluate sedimentation and erosion of the Orashi River in Nigeria. Also, Rahman et al. (2024) showed that the integration of geospatial methods and predictive models has high accuracy in identifying areas susceptible to riverbank erosion. Tzeppenlis et al. (2022) also presented an integrated system for monitoring coastal and riparian areas based on remote sensing and machine learning, which had high efficiency in coastal zone management.

In Iran, the use of satellite imagery and GIS in river geomorphology studies has expanded in recent years. Studies by Ilanloo (2011, 2020, 2025) have demonstrated the role of climatic, tectonic, and land use change factors in the evolution of geomorphological forms and river system dynamics. Also, Ebrahimi et al. (2025) investigated the changes in alluvial and coastal plains of the Bahmanshir River over the past

five decades and showed the role of human activities in the expansion of alluvial sediments and changes in the morphological pattern of the region. Furthermore, Soleimani and Ilanloo (2025) used satellite imagery to analyze the trend of Bahmanshir River bed changes, and Ilanloo and Karam (2020) emphasized the importance of the Morphological Quality Index (MQI) in assessing the hydromorphological status of rivers.

Despite significant advances in the application of remote sensing and GIS, the complexity of relationships between hydrological, climatic, geomorphological, and human factors has made it necessary to use advanced analytical methods. Among these, machine learning algorithms, due to their ability to model nonlinear relationships, increase prediction accuracy, and analyze big data, are recognized as one of the most important modern tools in studies of erosion, sedimentation, and river morphological changes, and have found wide application in geomorphology studies in recent years.

The present study, aiming to monitor riverbank erosion in the Sefidroud watershed using a machine learning approach, has fundamental innovations and differences compared to previous studies. While previous studies such as Nguyen et al. (2026) used 4 Boosting models in the Mekong Delta of Vietnam, and Prakash et al. (2024) applied 3 models in Assam State, India, and Tzeppenlis et al. (2022) merely compared 2 Deep Learning models in Thermaikos Gulf, Greece, the present study, compares 4 different models including Random Forest, XGBoost, LightGBM and CatBoost simultaneously and evaluates their performance with 5 comprehensive criteria including Accuracy, Precision, Recall, F1-Score, and ROC-AUC. Also, unlike studies that have only used raster data (Sajedi-Hosseini et al., 2020), this study has utilized a combination of raster and tabular data.

The Sefidroud watershed, as the largest Iranian delta on the southern shores of the Caspian Sea, is considered one of the most important sedimentary and geomorphological units in the country. This delta covers a large part of the Gilan plain and, due to the presence of diverse coastal, riverine, and marine deposits, has high geomorphological dynamics. In addition to natural processes, including sedimentation, erosion, Caspian Sea level fluctuations, and river flow regime changes, human activities such as the expansion of agricultural lands, urban development, construction of hydraulic structures, and land use changes also play an important role in the morphological evolution of this region. The high population concentration, economic dependence of the region on natural resources, and infrastructure development have doubled the importance of investigating morphological changes and erosion trends in this watershed. Therefore, accurate recognition of sedimentation and erosion patterns and analysis of morphological changes can provide a suitable basis for implementing development projects, management planning, protection of fertile lands, and achieving sustainable development in the Sefidroud Delta.

2. Materials and Methods

2.1 Study Area Introduction

The Sefidroud River is one of the most important rivers in northern Iran, originating from the mountains of Kurdistan and, after the confluence of the Qezel-Ozan and Shahroud rivers in the southern part of Rudbar County in Gilan Province, continues its course toward the Caspian Sea. This river, with a length of approximately 765 kilometers, after passing through the provinces of Zanjan, Qazvin, and Gilan, finally flows into the Caspian Sea near Bandar-e Kiashahr. The study area starts from the Sefidroud Dam reservoir and ends at the river's entry point into the Caspian Sea. The geographical location of the Sefidroud area is between 49°29' to 90°20' east longitude and 16°46' to 10°18' north latitude (Fig. 1). In addition to its hydrological importance, the Sefidroud plays a fundamental role in supplying agricultural water, recharging groundwater, and developing human settlements in the region.

The Sefidroud River, with its extensive and complex watershed, plays a vital role in supplying agricultural water, recharging groundwater aquifers, and supporting the development of settlements and local economic activities. The average discharge of the river varies in different parts of the basin; at the point where the river exits Zanjan Province into Qazvin, the average discharge is reported to be about 45 cubic meters per second, and this amount increases to about 300 to 400 cubic meters per second in the lower sections near the river mouth, due to the collection of lateral flows and tributaries. This significant water flow has made the Sefidroud known as the main source of water and energy for various uses in northern Iran. (Fig. 1).

2.2 Research framework

This research was conducted with the aim of monitoring and predicting soil erosion in the Sefidroud watershed based on the integration of remote sensing data, Geographic Information Systems (GIS), and machine learning algorithms. The overall research process consisted of four main stages: (1) preparation and preprocessing of spatial data, (2) extraction of effective factors influencing erosion, (3) training and evaluation of machine learning models, and (4) generation of the final erosion susceptibility map (Fig. 2).

2.3 Spatial Database Preparation

To model erosion spatially in the Sefidroud watershed, a multi-source spatial database was prepared by integrating remote sensing data, topographic data, geological information, soil, climate, and hydrology. All data were processed in a GIS environment and, after standardizing the coordinate system, spatial cell size, and study area extent, were prepared as raster layers with a spatial resolution of 30 meters. All layers were converted to the UTM Zone 39N (WGS84) coordinate system and clipped to the boundary of the Sefidroud watershed to enable spatial analysis and feature extraction within a unified spatial framework. To eliminate spatial scale discrepancies, all layers were resampled, and data containing invalid values (NoData) were removed or corrected. This step enhanced data compatibility and reduced errors arising from the integration of different information sources.

2.4 Extraction of Predictor Variables

Based on previous studies and considering the geomorphological characteristics of the Sefidroud watershed, fourteen environmental variables were selected as controlling factors of erosion. These variables represent topographic, hydrological, geological, climatic, and land surface conditions and play an effective role in runoff generation, flow concentration, sediment transport, and erosion occurrence.

a) Remote Sensing Variables

Sentinel-2 and Landsat satellite images, after geometric and atmospheric correction, were used to extract land surface information. From these images, the NDVI index was extracted as an indicator of vegetation cover status. This index reflects the density and health of vegetation and is considered one of the most important factors in reducing erosion by increasing soil surface protection and reducing rainfall droplet energy.

Additionally, the land cover map was extracted from satellite image classification, and different land cover classes were used as one of the input variables to the model.

b) Topographic Variables

Topographic features were extracted from the Digital Elevation Model (DEM). These variables included:

- Elevation
- Slope
- Aspect
- Curvature
- Stream Power Index (SPI)

These variables are the primary controllers of runoff concentration, flow velocity, erosive energy, and sediment transport across the watershed and are recognized as the most important environmental factors in most erosion prediction models.

c) Geological and Soil Variables

Lithology information was extracted from geological maps, and the various rock formations of the area were incorporated into the model as raster data.

Additionally, the soil type map was prepared from soil survey information. Geological characteristics and soil type are among the most important controlling factors of erosion processes due to their direct influence on material resistance to erosion, permeability, and sediment production.

d) Hydrological and Climatic Variables

To represent the hydrological conditions of the watershed, four variables were incorporated into the model:

- River Distance
- Discharge
- Annual Precipitation
- Flow Accumulation

Precipitation is the most important factor initiating water erosion and controls the intensity and amount of runoff generated. On the other hand, distance from the river and flow discharge indicate the sediment transport capacity and the likelihood of bank erosion occurring in different areas of the watershed.

2.5 Formation of the Training Database

After preparing all environmental layers, a spatial database consisting of fourteen independent variables and one dependent variable was established. The dependent variable represented the occurrence or non-occurrence of erosion and was determined through field surveys, satellite image interpretation, and available information.

A total of 1000 samples were selected for model development, including 499 samples with erosion and 501 samples without erosion. The values of all environmental variables were extracted for each sample and organized in the form of a data matrix for training the machine learning algorithms.

Finally, after standardization, the data were split into 80% for training and 20% for testing. Additionally, to mitigate the effect of class imbalance, the Class Weighting method was applied during the training process to improve the models' performance in detecting both classes.

3.Results

3.1 Changes in environmental indicators

In order to evaluate the capability of machine learning algorithms in predicting erosion, four models, namely XGBoost, Random Forest, CatBoost, and LightGBM, were developed, and their performances were evaluated using the Receiver Operating Characteristic (ROC) curve. The results indicated that all four models possess a desirable ability to distinguish between erodible and non-erodible areas; however, there are differences in their discriminative power.

Based on the ROC curve (Fig. 3), the Random Forest model, with an AUC value of 0.962, provided the best performance among the investigated models. Subsequently, the LightGBM (AUC = 0.947) and XGBoost (AUC = 0.943) models ranked next, while CatBoost, with an AUC value of 0.902, showed the lowest performance among the four models. The proximity of the ROC curve of the Random Forest model to the upper-left corner of the plot indicates the high capability of this model in correctly discriminating between non-erosion and erosion classes. Furthermore, the slight difference between the XGBoost and LightGBM models suggests that these two algorithms also possessed suitable predictive capabilities, although their performance was slightly lower than that of the Random Forest model.

Overall, the evaluation results demonstrate that tree-based models, particularly Random Forest, have been able to model the nonlinear relationships between environmental factors and the occurrence of erosion with greater accuracy compared to other algorithms (Fig. 3).

Table 1 shows the priority of importance of the variables used in the models used in the research. In the LightGBM model, the Stream Power Index (SPI), with an importance score of 307, had the greatest contribution to erosion prediction, followed by lithology (299), precipitation (297), and flow accumulation (268). In addition, NDVI, distance to the river, elevation, discharge, and land cover also contributed to improving model performance. Although the feature importance values in LightGBM are calculated based on the number of tree splits and are therefore not directly comparable with those of the other models, the ranking of variables clearly indicates that hydrological and topographic factors play dominant roles in the erosion process.

Similarly, in the CatBoost model, precipitation, with an importance score of 16.169, was identified as the most influential variable, followed by flow accumulation, SPI, lithology, and NDVI. In addition, distance to the river, slope aspect, surface curvature, and soil type, although contributing to a lesser extent, also played a role in improving model accuracy.

Overall, the comparison of the four machine learning models demonstrates that precipitation, flow accumulation, the Stream Power Index (SPI), lithology, and the Normalized Difference Vegetation Index (NDVI) consistently ranked among the most influential variables controlling riverbank erosion. This agreement among the different algorithms enhances the reliability of the results and indicates that riverbank erosion within the Sefidroud Watershed is primarily governed by climatic, hydrological, topographic, and land surface characteristics. From a watershed management perspective, these findings suggest that runoff and rainfall management, control of flow concentration, conservation of vegetation cover, consideration of lithological characteristics, and continuous monitoring of areas with high stream power should be prioritized in riverbank protection and stabilization programs. The strong consistency among the different machine learning models further increases the robustness of the findings and provides a sound scientific basis for sustainable watershed management in the Sefidroud Watershed and other watersheds with similar environmental conditions.

Table 1
Priority of importance of variables used in the models

row	Variable	Random Forest	XGBoost	LightGBM	CatBoost
1	Flow Accumulation	0.117	0.135	268	15.942
2	Precipitation	0.126	0.110	297	16.169
3	SPI	0.109	0.102	307	15.641
4	NDVI	0.115	0.107	240	15.122
5	Litology	0.115	0.104	299	15.372
6	River Distance	0.047	0.047	185	2.704
7	Soil	0.046	0.047	—	2.422
8	Curvature	0.043	0.048	—	2.521
9	Slope	—	0.048	—	—
10	Landcover	—	—	151	—
11	Discharge	—	—	165	—
12	Aspect	—	—	—	2.561
13	Elevation	—	—	179	—

To evaluate the ability of the machine learning models to classify riverbank erosion into high- and low-risk areas, the performance of four algorithms, namely Random Forest, XGBoost, LightGBM, and CatBoost, was assessed using Accuracy, Precision, Recall, F1-score, and the confusion matrix (Table 2). The results showed that although all models achieved very high recall values, considerable differences were observed in the remaining evaluation metrics.

The Random Forest model achieved the highest overall accuracy (Accuracy = 0.915) among all evaluated models. It also produced the highest F1-score (0.910), indicating the best balance between precision and recall. Moreover, the model generated only two false negatives (FN = 2), demonstrating its ability to correctly identify nearly all high-erosion areas. In addition, Random Forest correctly classified 3,195 true negatives (TN), the highest value among all models, highlighting its superior capability in identifying low-risk areas. Although the number of false positives (FP = 2,192) remained relatively high, it was substantially lower than that of the other algorithms. Consequently, Random Forest achieved the best trade-off between accurately detecting high-risk areas and minimizing false alarms. Therefore, this algorithm can be considered the most suitable model for practical watershed management and riverbank erosion hazard mapping in the Sefidroud Watershed.

Although XGBoost achieved a similarly high Recall (0.995), its overall performance was inferior to that of Random Forest. The model also produced only two false negatives, indicating a high capability for

detecting high-risk erosion areas. However, the large number of false positives (FP = 4,379) substantially reduced its Precision (0.085) and consequently its overall Accuracy (0.244). These findings indicate that XGBoost has a strong tendency to classify samples as high-risk, thereby minimizing the probability of missing critical erosion areas at the expense of generating a large number of false alarms. Such behavior limits its practical applicability in hazard mapping and management.

The LightGBM model achieved a perfect Recall of 1.000, successfully identifying all high-risk erosion areas without producing any false negatives (FN = 0). This characteristic represents a clear advantage in applications where missing hazardous locations is unacceptable. Nevertheless, the model generated 4,427 false positives, resulting in a substantial reduction in Precision (0.084) and Accuracy (0.236). Therefore, despite its excellent sensitivity, the excessive number of false alarms considerably reduces its reliability for practical watershed management and decision-making.

Among the four algorithms, CatBoost exhibited the weakest overall performance. Similar to LightGBM, it achieved a perfect Recall (1.000) and correctly identified all high-risk erosion areas. However, the generation of 5,050 false positives led to markedly lower Precision (0.075), Accuracy (0.128), and F1-score (0.139). These results indicate that CatBoost has a pronounced tendency to assign samples to the high-risk class, causing a large proportion of low-risk areas to be incorrectly classified as hazardous. Such performance substantially limits its usefulness for practical applications, particularly in erosion hazard mapping.

Table 2
Comparison of the performance of machine learning models based on the confusion matrix and classification evaluation indices.

Model	Accuracy	Precision	Recall	F1-Score	TN	FP	FN	TP
XGBoost	0.244	0.085	0.995	0.156	1008	4379	2	405
Random Forest	0.915	0.935	0.995	0.910	3195	2192	2	405
CatBoost	0.128	0.075	1.000	0.139	337	5050	0	407
LightGBM	0.236	0.084	1.000	0.155	960	4427	0	407

TN: True Negative, FP: False Positive, FN: False Negative, TP: True Positive.

The weighted confusion matrices (Fig. 4) show that the Random Forest model performed best among the four algorithms with an accuracy of 62.1% and an F1-Score of 0.270. This model achieved the best balance between sensitivity and precision, identifying 405 out of 407 high-risk points (Recall = 99.5%) and producing 2192 false positives (FP). In contrast, although LightGBM and CatBoost identified all high-risk points (Recall = 100%), the very high number of false positives (4427 and 5050, respectively) limited their practical application. These results emphasize that in environmental problems with unbalanced data, simultaneous evaluation of multiple indicators (especially F1-Score and number of false positives) is essential for selecting the appropriate model.

To investigate the relationships among the variables affecting riverbank erosion and to assess the degree of interdependence among them, the Spearman rank correlation matrix was employed (Fig. 5). The results indicated that most independent variables exhibit weak to moderate correlations with one another, suggesting their relative independence and the suitability of the dataset for machine learning modeling.

However, among the examined variables, the highest correlation coefficient was observed between the Stream Power Index (SPI) and Flow Accumulation, with a value of 0.97, indicating a very strong positive relationship between these two hydrological indices. This finding demonstrates that both variables provide similar information regarding surface runoff concentration and flow conveyance capacity across the watershed.

Following these two variables, the highest correlation was observed between Slope and Flow Accumulation, with a coefficient of 0.50, indicating a moderate positive correlation. Additionally, a positive correlation of 0.39 was found between Slope and SPI, suggesting that flow capacity increases in steeper areas. These relationships are consistent with the hydrological and geomorphological principles governing runoff concentration and sediment transport in mountainous environments.

The results also revealed a negative correlation between Elevation and Precipitation, with a coefficient of -0.36. This finding indicates that within the study area, precipitation decreases with increasing elevation. Furthermore, the negative correlation between Elevation and Distance to River (-0.26) suggests that low-elevation areas are generally located in proximity to the drainage network.

On the other hand, a positive correlation of 0.31 was observed between Precipitation and Distance to River, indicating a relative association between precipitation patterns and the spatial distribution of the river network. Moreover, Precipitation and NDVI exhibited a positive correlation of 0.25, suggesting that increased precipitation leads to improved vegetation cover. In contrast, the correlation between NDVI and River Width was negative and very weak (-0.19), indicating a limited relationship between these two variables in the study area.

Overall, the Spearman correlation matrix demonstrated that, except for the very strong relationship between SPI and Flow Accumulation, the other environmental variables exhibit correlation coefficients lower than 0.50. This finding indicates that most of the selected variables provide relatively independent information regarding the climatic, hydrological, topographic, and geomorphological conditions of the Sefidroud watershed. Consequently, the chosen set of variables possesses sufficient information diversity for modeling riverbank erosion.

To assess the reliability of the Random Forest model predictions, uncertainty analysis was performed using the nonparametric bootstrap method with 100 replicates, and 95% prediction intervals were calculated for each sample (Fig. 6). In this Fig, the blue line represents the mean of the predicted values and the blue shaded area represents the range of the 95% prediction intervals. The samples are sorted by predicted values, and the vertical axis represents the predicted erosion values on a logarithmic scale.

The results showed that for samples with low erosion values (log values less than about 0.2), the prediction intervals are relatively narrow, indicating the stability of the predictions and the low uncertainty of the model in this range. This indicates that the model has high reliability in predicting areas with low erosion risk and that the variability of the predictions across different bootstrap iterations is limited.

In contrast, as the predicted values of erosion increase, especially at logarithmic values greater than about 0.5, the width of the prediction intervals increases significantly. This pattern indicates an increase in the uncertainty of the model in predicting areas with severe erosion. The increase in the width of the prediction intervals at larger values indicates an increase in the dispersion of the predictions in the bootstrap samples and indicates that the model has less confidence in severe erosion conditions. Such behavior is common in modeling environmental and geomorphological processes, because extreme events are usually less frequent in the training data set and are therefore associated with greater variability.

Figure 7 shows the SHAP Summary Plot for the Random Forest model in predicting riverbank erosion susceptibility in the Sefidrud watershed. In this plot, the variables are ordered from top to bottom based on the absolute mean of the SHAP values; therefore, the variables at the top of the plot have the largest contribution to the model predictions. The horizontal axis shows the SHAP value, which indicates the extent to which each variable affects the model output; positive SHAP values increase the probability of erosion prediction, and negative values decrease it. Also, the color of the points indicates the actual value of each variable; red indicates high values and blue indicates low values of that feature.

Based on the results of this analysis, precipitation is the most important factor in predicting riverbank erosion susceptibility. High precipitation values are mainly associated with positive SHAP values, indicating that increased precipitation is associated with an increase in the model's predicted probability of erosion. This pattern is consistent with the hydrological role of precipitation in increasing surface runoff, stream discharge, and shear stress on riverbanks. Elevation is the second most important variable in the model, with lower elevations mostly associated with positive SHAP values and higher elevations with negative SHAP values. This result indicates that low-lying areas, especially floodplains and downstream areas, are more susceptible to riverbank erosion than high-lying areas.

Lithology and soil type are also influential variables in the model. The distribution of SHAP values shows that some lithological units and some soil groups have increased erosion sensitivity, while others have a reducing role. This indicates the importance of geological and soil characteristics in controlling the stability of riverbanks and their resistance to erosion. Also, distance from the river has a relatively significant effect on the model output; So that points closer to the river are generally associated with positive SHAP values, indicating greater sensitivity of these areas to erosion, due to the direct impact of water flow and river processes.

The Hotspot Analysis map (Getis-Ord Gi) illustrates the spatial pattern of riverbank erosion susceptibility within the Sefidroud watershed. In this map, the positive and high values of the Gi statistic, displayed in

the yellow, orange, and red color spectrum, indicate the formation of statistically significant clusters of areas with high erosion susceptibility, while the blue colors represent cold spots or areas with low erosion susceptibility. The results show that the highest concentration of hot spots is located in the northeastern parts of the study area, where G_i^* values exceed 200, indicating the presence of highly significant clusters of erosion-prone zones. This suggests that the factors influencing erosion—such as slope, geomorphological characteristics, stream power, sediment type, land use, and vegetation cover—are acting simultaneously in this part of the watershed, creating favorable conditions for the intensification of riverbank erosion.

In contrast, the central, southern, and southwestern parts of the watershed are predominantly displayed in blue, which indicates negative or low G_i^* statistic values and the presence of cold spots. This condition shows that these areas are statistically less susceptible to erosion and possess more stable environmental and hydromorphological conditions. Furthermore, the light-colored and white areas located between the hot and cold spots represent zones with values close to the mean, which do not exhibit statistically significant clustering and can be considered transitional zones between stable and critical areas (Fig. 8.).

Figure 9 shows the riverbank erosion susceptibility zoning map. The riverbank erosion susceptibility zoning map shows that the distribution of erosion risk across the Sefidrud catchment area is not uniform and follows a clustered spatial pattern. The highest concentration of “high” and “very high” classes is observed along the main river course, indicating the dominant role of hydrological processes and flow dynamics in controlling bank erosion. This spatial pattern indicates that areas adjacent to the main channel are most susceptible to erosion due to the direct impact of water flow, runoff concentration, and shear stress on the riverbanks.

The northern part of the basin, especially the downstream area and close to the Sefidrud delta, has the highest density of the “very high” class. This situation can be attributed to the simultaneous effect of several geomorphological and hydrological factors; Including the reduction of the longitudinal slope of the river, the entry of the flow into the alluvial plain, the increase in the width of the channel, the development of meanders, the presence of alluvial sediments with low mechanical resistance and the concentration of the flow in this area. Such conditions increase the instability of the river banks and intensify erosion and are consistent with the behavioral pattern of large rivers leading to deltas and coastal areas.

In the middle part of the basin, relatively continuous bands of the “high” and “very high” classes are seen on both sides of the river course, indicating the direct impact of the morphological characteristics of the channel on the occurrence of erosion. In this part, factors such as the radius of curvature of the bends, the flow power index (SPI), flow accumulation, distance from the river and the lithological and sedimentary characteristics of the bed play an important role in increasing the susceptibility to erosion. The presence of these high-risk strips along the main channel indicates that local changes in river geometry and hydraulic conditions play a decisive role in the spatial distribution of erosion.

In contrast, sections far from the main river course are mainly in the “very low” class, indicating a significant decrease in the probability of riverbank erosion in these areas. With increasing distance from the active channel, the direct effect of water flow, hydraulic energy, and bank erosion decreases, resulting in an increase in the relative stability of the land. In addition, the presence of scattered patches with “low” and “medium” classes in different parts of the basin indicates that some local factors such as land use change, reduced vegetation density, specific geological features, or local topographic conditions can also provide the basis for increased susceptibility to erosion even at greater distances from the river.

The spatial pattern observed in this map is also fully consistent with the results of the Random Forest model variable importance analysis. Based on the model results, variables such as precipitation, flow accumulation, flow power index (SPI), lithology, and vegetation cover index (NDVI) are the most important factors affecting the prediction of riverbank erosion.

4. Discussion

4.1. Performance Analysis of Machine Learning Models

The results of the study showed that the Random Forest model, with an accuracy of 93.48%, a precision of 91.50%, a sensitivity of 88.66%, an F1-score of 91.01%, and an area under the ROC curve of 98.38%, had the best performance among the four algorithms examined. This superior performance can be attributed to the structural characteristics of the Random Forest model, which include resistance to data noise, the ability to simultaneously handle quantitative and qualitative variables, the ability to model nonlinear relationships, and the reduction of overfitting through the ensemble of a large number of decision trees. These findings are consistent with the results of previous studies (Avand et al., 2023; Chakraborty et al., 2020), which have introduced Random Forest as one of the most accurate machine learning algorithms for environmental hazard modeling.

The XGBoost, LightGBM, and CatBoost models also showed acceptable performance; however, due to the high number of false positives, which were 4,379, 4,427, and 5,050 cases, respectively, their operational efficiency in erosion susceptibility mapping is limited. This indicates that although boosting-based models have a high ability to identify susceptible areas (with a sensitivity of 100% for LightGBM and CatBoost), they have a greater tendency to classify samples as susceptible, which can lead to reduced management efficiency and false alarms. This behavior has also been observed in previous studies (Harrison et al., 2018).

Comparison of the models based on the weighted confusion matrix showed that Random Forest established the best balance between accuracy and sensitivity by identifying 405 out of 407 susceptible points (sensitivity = 99.5%) and producing 2,192 false positives.

4.2. Analysis of Variable Importance Using SHAP

The analysis of variable importance using the SHAP method showed that precipitation had the greatest contribution to predicting susceptibility to riverbank erosion, which is consistent with the results of Nguyen et al. (2026). It was followed, in order, by flow accumulation, the Stream Power Index (SPI), lithology, and the Normalized Difference Vegetation Index (NDVI). This finding is consistent with the prevailing geomorphological and hydrological principles governing riverine erosion.

The high importance of hydrological indices such as the Stream Power Index (SPI) and flow accumulation indicates that processes related to flow concentration and runoff generation play an essential role in intensifying riverbank erosion. The Stream Power Index, which is a combination of flow accumulation and slope, indicates the ability of flow to erode the riverbed and riverbanks and transport sediment.

The importance of lithology in the model indicates the role of the different resistance of geological units to erosion, which is consistent with the results of Eilanlou (2011, 2020), who confirmed the role of geological factors in the evolution of geomorphological forms.

The role of the NDVI as one of the important variables indicates the effect of vegetation cover on reducing erosion. Vegetation cover reduces erosion by increasing soil cohesion, reducing runoff velocity, and protecting riverbanks, thereby reducing the probability of erosion.

4.3. Analysis of the Spatial Patterns of Erosion

The erosion susceptibility zoning map shows that the distribution of erosion susceptibility across the Sefidrud Basin is not uniform and follows a clustered pattern. The highest susceptibility, particularly the very high and high classes, was observed along the main river channel, indicating the dominant role of flow dynamics and hydrological processes in controlling riverbank erosion. This spatial pattern is also consistent with the Getis-Ord G_i^* hot spot analysis, such that the highest concentration of hot spots was observed in the northeastern parts of the study area, with G_i^* values greater than 200.

The northern part of the basin, particularly the downstream area near the Sefidrud Delta, has the largest area of the "very high" class. This condition can be attributed to the simultaneous influence of several hydrological and geomorphological factors, including the entry of the flow into the alluvial plain, a decrease in the longitudinal slope of the river, the development of meanders, an increase in channel width, the concentration of flow in this area, and the presence of alluvial sediments with low mechanical resistance. These conditions cause increased instability of the riverbanks and intensified erosion and are consistent with the behavioral pattern of large rivers flowing into deltas and coastal areas.

In the central part of the basin, relatively continuous belts of the "very high" and "high" classes are observed on both sides of the river channel, indicating the direct influence of the morphological characteristics of the channel on the occurrence of erosion. In this section, factors such as flow accumulation, stream power, the radius of curvature of bends, distance from the river, and the sedimentary and lithological characteristics of the riverbed play an important role in increasing erosion susceptibility.

In contrast, areas far from the main river channel are mainly classified as "very low," indicating a significant reduction in the probability of riverbank erosion in these areas. As the distance from the riverbed increases, the direct effects of water flow, hydraulic energy, and riverbank erosion decrease, and consequently the relative stability of the land increases. However, the presence of scattered patches of the "low" and "moderate" classes in different parts of the basin indicates the role of some local factors, such as reduced vegetation density, land-use change, specific geological characteristics, or topographic conditions, in increasing erosion risk even at greater distances from the river.

4.4. Analysis of Model Uncertainty

The uncertainty analysis using the bootstrap method with 100 repetitions showed that, for samples with low erosion values (logarithmic values below approximately 0.2), the prediction intervals are relatively narrow, indicating the stability of the predictions and low model uncertainty in this range. This indicates the high reliability of the model in predicting areas with low erosion risk, which is consistent with the results of the study by Folharini et al. (2022).

5. Conclusions

The results of this study showed that the Random Forest model, after optimizing the parameters using Grid Search, provided the best performance in predicting the sensitivity of riverbank erosion in the Sefidrood basin, with Accuracy values of 91.50%, Precision of 93.48%, Recall of 88.66%, F1-Score of 91.01%, and ROC-AUC of 98.38%. These results indicate that hyperparameter optimization plays an important role in increasing the predictive power of the model and creates a suitable balance between Precision and Recall. Previous studies have also reported that appropriate adjustment of Random Forest parameters can significantly increase the accuracy of the model in environmental hazard modeling problems; therefore, the results of this study are consistent with the findings of Nguyen et al. (2026) and Rahman et al. (2024).

The superiority of Random Forest over other algorithms under study can be attributed to its ability to model nonlinear relationships, resist data noise, simultaneously handle quantitative and qualitative variables, and reduce overfitting by aggregating a large number of decision trees. These features have led Random Forest to be introduced as one of the most accurate machine learning algorithms in many studies related to modeling erosion, landslides, floods, and other geomorphological hazards. The results of this study also confirm this and show that even in the presence of multiple variables and complex relationships between environmental factors, this model was able to identify the spatial pattern of erosion with high accuracy.

Analysis of the importance of variables using the SHAP method showed that precipitation, flow accumulation, stream power index (SPI), lithology, and vegetation cover index (NDVI) have the greatest contribution to predicting riverbank erosion susceptibility. The high importance of precipitation and hydrological indices indicates that processes related to runoff generation and flow concentration play a fundamental role in intensifying bank erosion. On the other hand, the importance of lithology indicates

that the different resistance of geological units to erosion is an important controller in the stability of river banks. Also, the significant role of NDVI indicates that vegetation cover reduces the probability of erosion by reducing runoff velocity, increasing soil cohesion and protecting river banks. These findings are consistent with the results of studies Acquaotta et al. (2019), Jonathan & Charles (2025) and Acquaotta et al. (2019) that reported the role of hydrological factors, vegetation cover and geological features in controlling river bank erosion.

The generated sensitivity map also shows a logical spatial pattern consistent with the geomorphological conditions of the basin. The highest sensitivity was observed along the main course of the river and especially in the downstream and northern parts of the basin. This pattern can be explained by the decrease in slope, increase in channel width, development of meanders, presence of loose alluvial deposits and concentration of flow in these areas. In contrast, the sections far from the active channel were mainly in low sensitivity classes, indicating a decrease in the direct effect of hydraulic processes on bank erosion. This spatial pattern is consistent with studies conducted in similar basins and shows that the developed model was able to identify critical areas with appropriate accuracy.

One of the strengths of this study is the use of SHAP analysis to interpret the model. Many studies have focused solely on providing accuracy indices, while in this study, in addition to evaluating model performance, the contribution of each variable in model decision-making was also examined. This increased the interpretability of the model and the reliability of the results and can help water resource managers in identifying the main factors controlling erosion.

Also, the use of Bootstrap-based uncertainty analysis allowed the stability of model performance to be evaluated. The results of this analysis showed that the changes in the evaluation indicators were limited and the model had good stability. This increases the generalizability of the results to similar conditions.

Despite the favorable performance of the model, this study also has limitations. First, the model was developed based on existing data and current conditions of the basin and future changes due to climate change or land use change were not included in it. Second, although the main factors affecting erosion were used in the model, the inclusion of variables such as instantaneous discharge, temporal changes in the flow regime, and hydrodynamic data can increase the prediction accuracy in future studies. Also, the use of spatial cross validation and multi-temporal data can further improve the generalizability of the model.

Declarations

Funding

The author declares that no funds, grants, or other financial support were received during the preparation of this manuscript.

Consent to Participate

Consent to Participate declaration: not applicable.

Consent to Publish

Consent to Publish declaration: not applicable.

Competing Interests

The author declares no competing interests.

Data Availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Author Contributions

The author solely conceived the study, performed the analyses, interpreted the results, and wrote the manuscript. The author read and approved the final manuscript.

References

1. Acquafredda, F., Faccini, F., Fratianni, S., Paliaga, G., Sacchini, A., & Vilímek, V. (2019). Increased flash flooding in Genoa metropolitan area: A combination of climate changes and soil consumption. *Meteorology and Atmospheric Physics*, 131(4), 1099–1110. <https://doi.org/10.1007/s00703-018-0623-0>
2. Avand, M., Mohammadi, M., Mircholi, F., Kaviani, A., & Tiefenbacher, J. P. (2023). A new approach for smart soil erosion modeling: Integration of empirical and machine-learning models. *Environmental Modeling & Assessment*, 28(1), 145–160. <https://doi.org/10.1007/s10666-022-09858-x>
3. Begam, S., & Sen, D. (2019). Mapping of moraine-dammed glacial lakes and assessment of their areal changes in the central and eastern Himalayas using satellite data. *Journal of Mountain Science*, 16(1), 77–94. <https://doi.org/10.1007/s11629-018-5023-1>
4. Chakraborty, R., Pal, S. C., Sahana, M., Mondal, A., Dou, J., Pham, B. T., & Yunus, A. P. (2020). Soil erosion potential hotspot zone identification using machine learning and statistical approaches in eastern India. *Natural Hazards*, 104(2), 1259–1294. <https://doi.org/10.1007/s11069-020-04213-3>
5. Ebrahimi, L., Ilanloo, M., & Hamidi, Z. (2025). Morphological changes in alluvial and coastal plains (Case example: Bahmanshir River). *Quantitative Geomorphological Research*, 13(4), 82–94. <https://doi.org/10.22034/gmpj.2025.485368.1528>
6. Folharini, S., Vieira, A., Bento-Gonçalves, A., Silva, S., Marques, T., & Novais, J. (2022). Soil erosion quantification using machine learning in sub-watersheds of northern Portugal. *Hydrology*, 10(1), 7. <https://doi.org/10.3390/hydrology10010007>

7. Gomez, J. A., Kamran-Pishhesari, A., & Sattarvand, J. (2025). Automated rill erosion detection in tailing dams using UAV imagery and machine learning. *Arabian Journal for Science and Engineering*, 50(9), 6711–6726. <https://doi.org/10.1007/s13369-024-09550-0>
8. Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E., & Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, 6, e4794. <https://peerj.com/articles/4794/>
9. Ilanloo, M. (2011). A comparative study of fuzzy logic approach for landslide susceptibility mapping using GIS: An experience of Karaj Dam Basin in Iran. *Procedia - Social and Behavioral Sciences*, 19, 668–676. <https://doi.org/10.1016/j.sbspro.2011.05.184>
10. Ilanloo, M. (2020). Investigation of active tectonics of Dalaki catchment using geomorphic indices and geomorphological evidences. *Geography*, 18(66), 168–185. <https://doi.org/20.1001.1.27172996.1399.18.3.10.9>
11. Ilanloo, M. (2025). Analysis of the dynamics of geomorphic continuity under the influence of climate and land use change in the Zarrineh Rud Basin, northwestern Iran. *Quantitative Geomorphological Research*, 14(3), 166–147. <https://doi.org/10.22034/gmpj.2025.532441.1568>
12. Ilanloo, M. (2026). Analyzing morphological changes in the Sefidrood River using satellite images and GIS (Case study: Sefidrood River, Gilan). *Journal of Geography and Planning*, 30(96). <https://doi.org/10.22034/gp.2025.67811.3421>
13. Ilanloo, M., & Karam, A. (2020). Assessment of the hydromorphological condition of the Jajrood River using the Morphological Quality Index (MQI). *Journal of Geographical Sciences*, 20(56), 35–52. <https://doi.org/10.29252/jgs.20.56.35>
14. Jonathan, L. E., & Charles, A. U. (2025). Shoreline erosion and accretion analysis of the Orashi River, Rivers State, Nigeria: A geospatial and machine learning approach. *Asian Journal of Geographical Research*, 8(2), 27–44. <https://doi.org/10.9734/AJGR/2025/v8i2260>
15. Nguyen, K. A., & Chen, W. (2021). DEM- and GIS-based analysis of soil erosion depth using machine learning. *ISPRS International Journal of Geo-Information*, 10(7), 452. <https://doi.org/10.3390/ijgi10070452>
16. Nguyen, K. A., Chen, W., Lin, B. S., & Seeboonruang, U. (2020). Using machine learning-based algorithms to analyze erosion rates of a watershed in northern Taiwan. *Sustainability*, 12(5), 2022. <https://doi.org/10.3390/su12052022>
17. Nguyen, T. G., Dang, D. K., Ngo, C. T., Nguyen, H. D., Pham, V. T., Bui, Q. T., ... & Van Luu, N. (2026). Prediction of riverbank erosion susceptibility areas of the Vietnam Mekong Delta using machine learning and remote sensing in the context of geographical studies. *Physical Geography*, 47(2), 105–129. <https://doi.org/10.1080/02723646.2025.2555849>
18. Rahman, M. N., Saleheen, M. M., El Fadili, H., & Sarker, M. N. I. (2024). Assessment of riverbank erosion and its prediction using geospatial and machine learning techniques. In *Modern Cartography Series* (Vol. 12, pp. 493–509). Academic Press. <https://doi.org/10.1016/B978-0-443-23890-1.00019-0>

19. Roy, S., & Chintalacheruvu, M. R. (2024). Google Earth Engine-based morphometric parameter evaluation and comparative analysis of soil erosion susceptibility using statistical and machine learning algorithms in large river basins. *Earth Science Informatics*, 17(1), 75–97. <https://doi.org/10.1007/s12145-023-01133-1>
20. Saha, S., Roy, J., Arabameri, A., Blaschke, T., & Tien Bui, D. (2020). Machine learning-based gully erosion susceptibility mapping: A case study of eastern India. *Sensors*, 20(5), 1313. <https://doi.org/10.3390/s20051313>
21. Soleimani, B., & Ilanloo, M. (2025). Temporal and spatial zoning of riverbed changes using satellite imagery in the ArcMap environment: A case study of the Bahmanshir River. *Hydrology*, 26, xx–xx. <https://doi.org/10.22034/hyd.2025.64643.1766>
22. Soleimani, B., Ilanloo, M., & Gharoubi, M. (2026). Flood simulation and flood potential mapping using the integrated WMS and HEC-RAS models (Case study: Savadkuh County). *Iranian Journal of the Environment*, 13(1), 1220–1240. <https://doi.org/10.22059/ije.2026.411871.1908>
23. Su, J., Tang, R., & Lin, H. (2024). Simulation and spatio-temporal analysis of soil erosion in the source region of the Yellow River using machine learning method. *Land*, 13(9), 1456. <https://doi.org/10.3390/land13091456>
24. Tzepkenlis, A., Grammalidis, N., Kontopoulos, C., et al. (2022). An integrated monitoring system for coastal and riparian areas based on remote sensing and machine learning. *Journal of Marine Science and Engineering*, 10(9), 1322. <https://doi.org/10.3390/jmse10091322>
25. Zeghmar, A., Mokhtari, E., & Marouf, N. (2024). A machine learning approach for RUSLE-based soil erosion modeling in Beni Haroun Dam watershed, northeast Algeria. *Earth Science Informatics*, 17(4), 2921–2936. <https://doi.org/10.1007/s12145-024-01305-7>

Figures

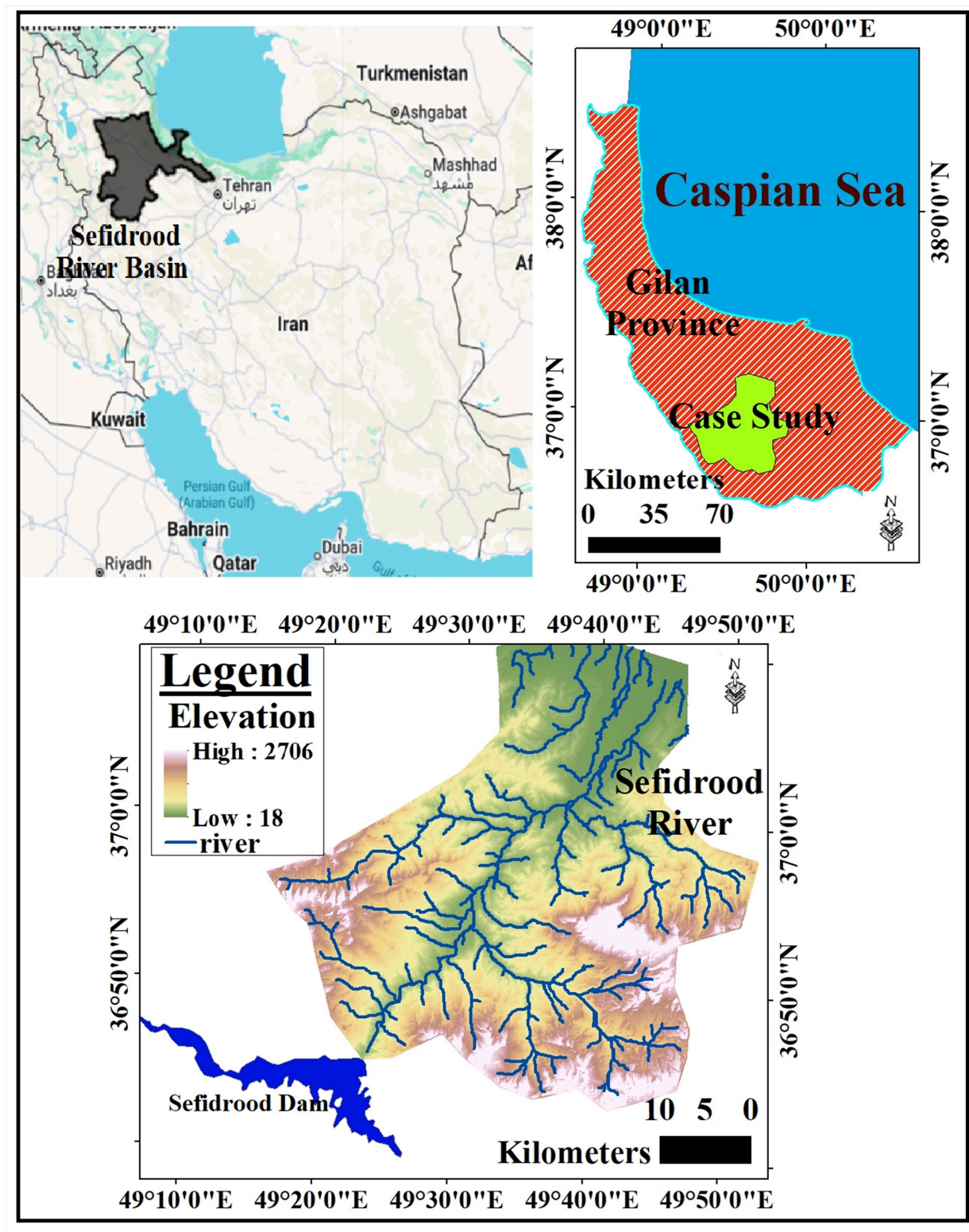


Figure 1

Location of the study area

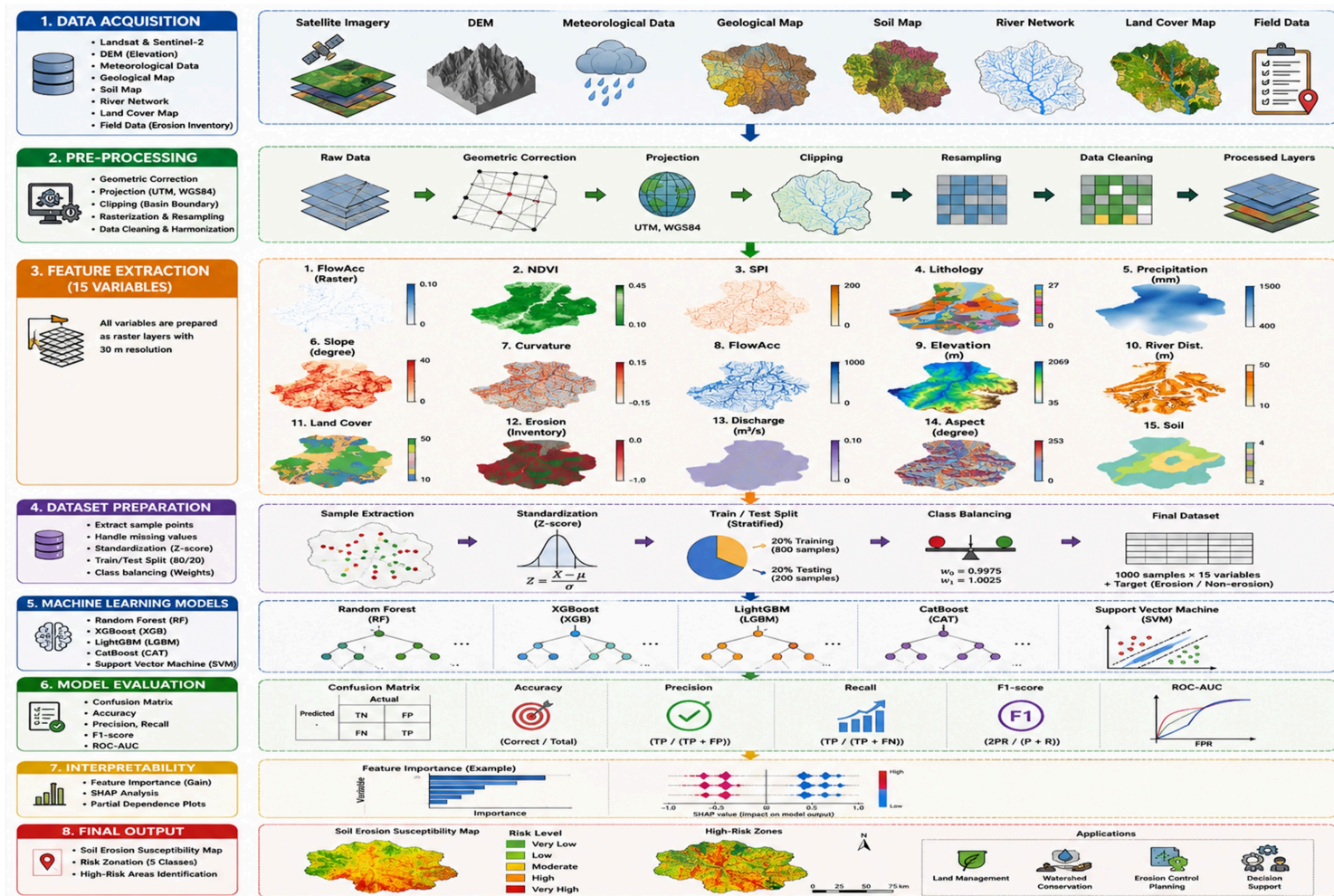


Figure 2

Research flowchart

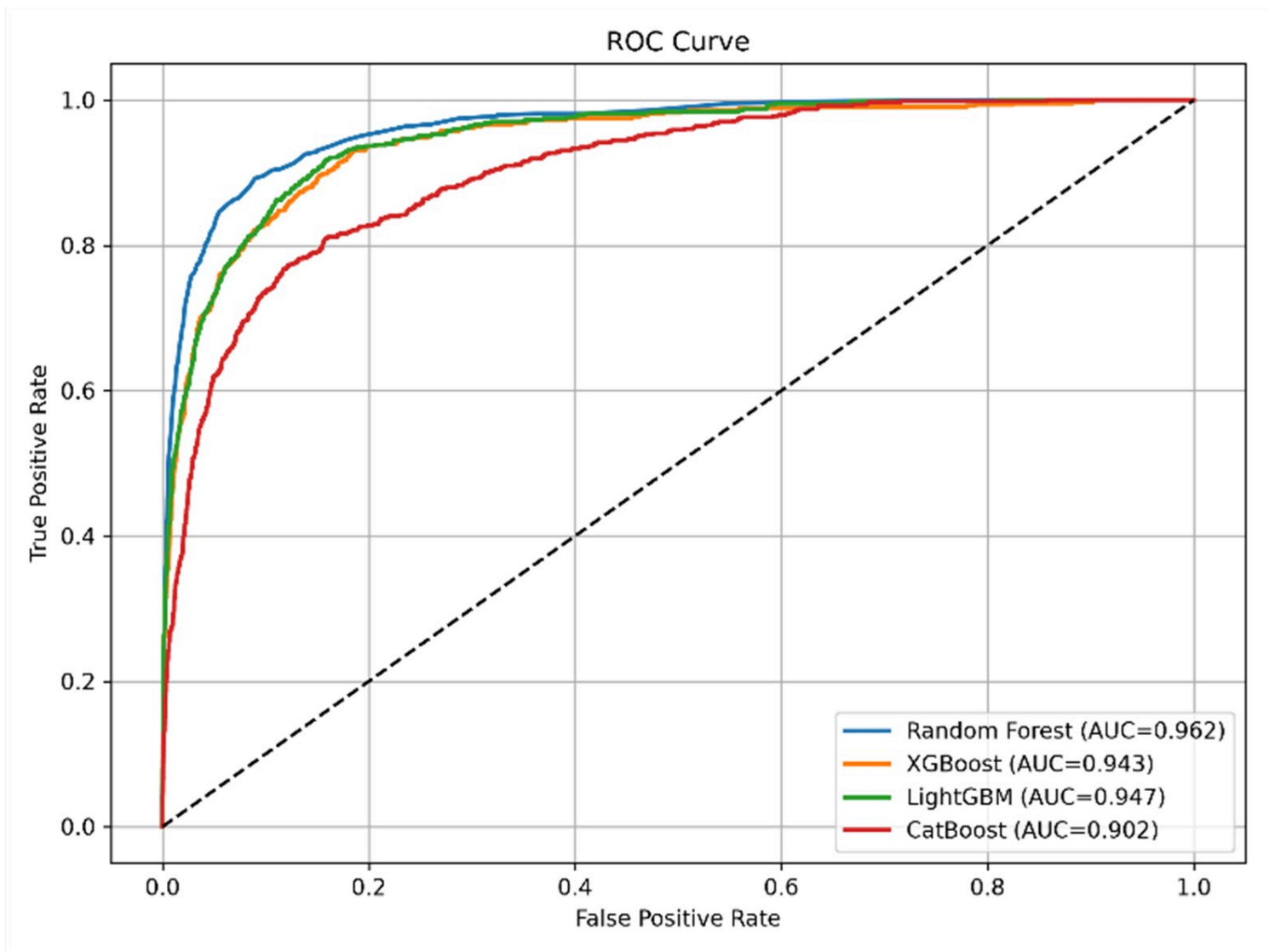


Figure 3

Comparison of the ROC curves of XGBoost, Random Forest, CatBoost models and LightGBM

Confusion Matrices (High Risk > 90th Percentile, Weighted)

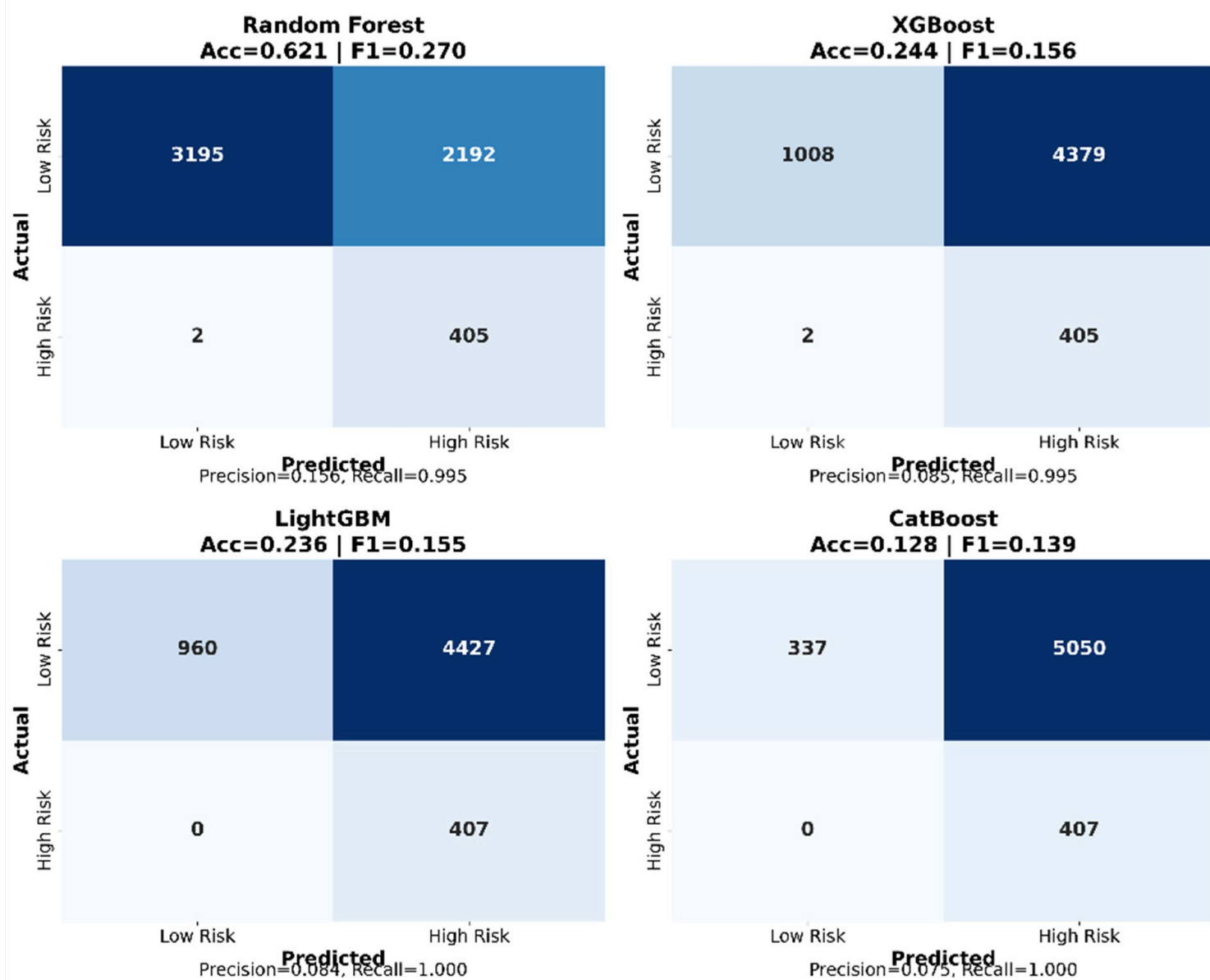


Figure 4

Confusion matrix and classification evaluation indices

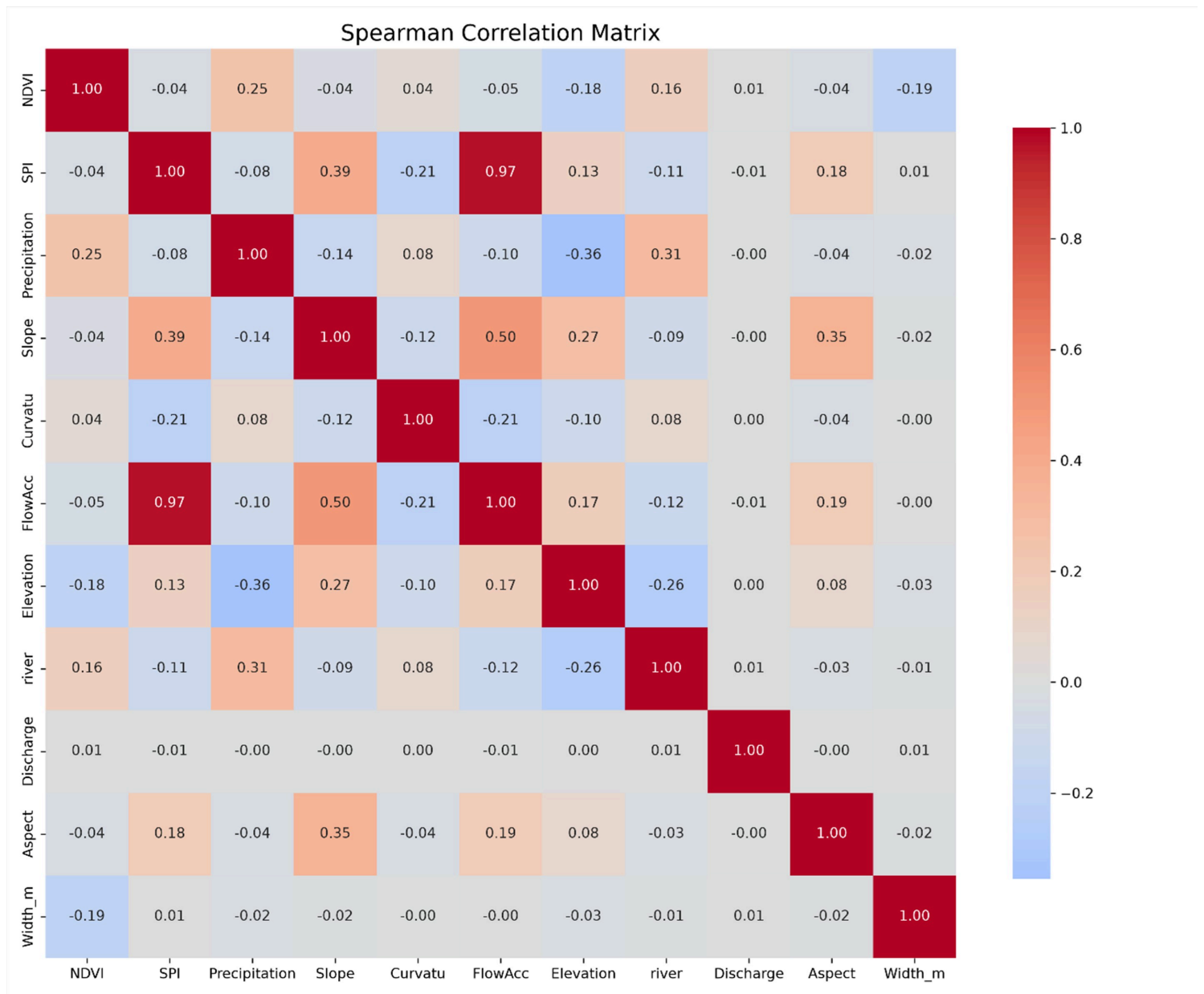


Figure 5

Spearman correlation of variables

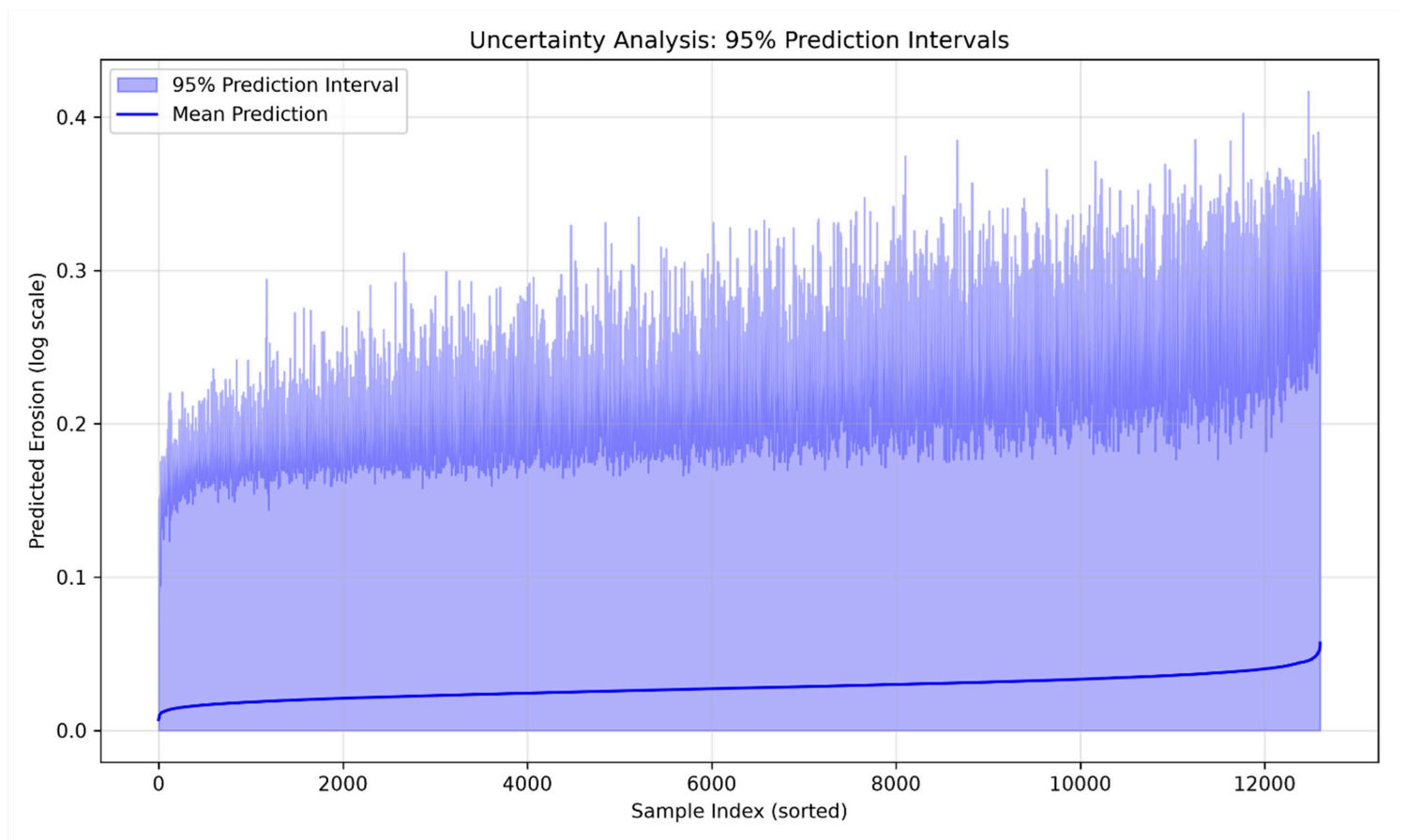


Figure 6

Uncertainty analysis: 95% prediction intervals for the Random Forest model

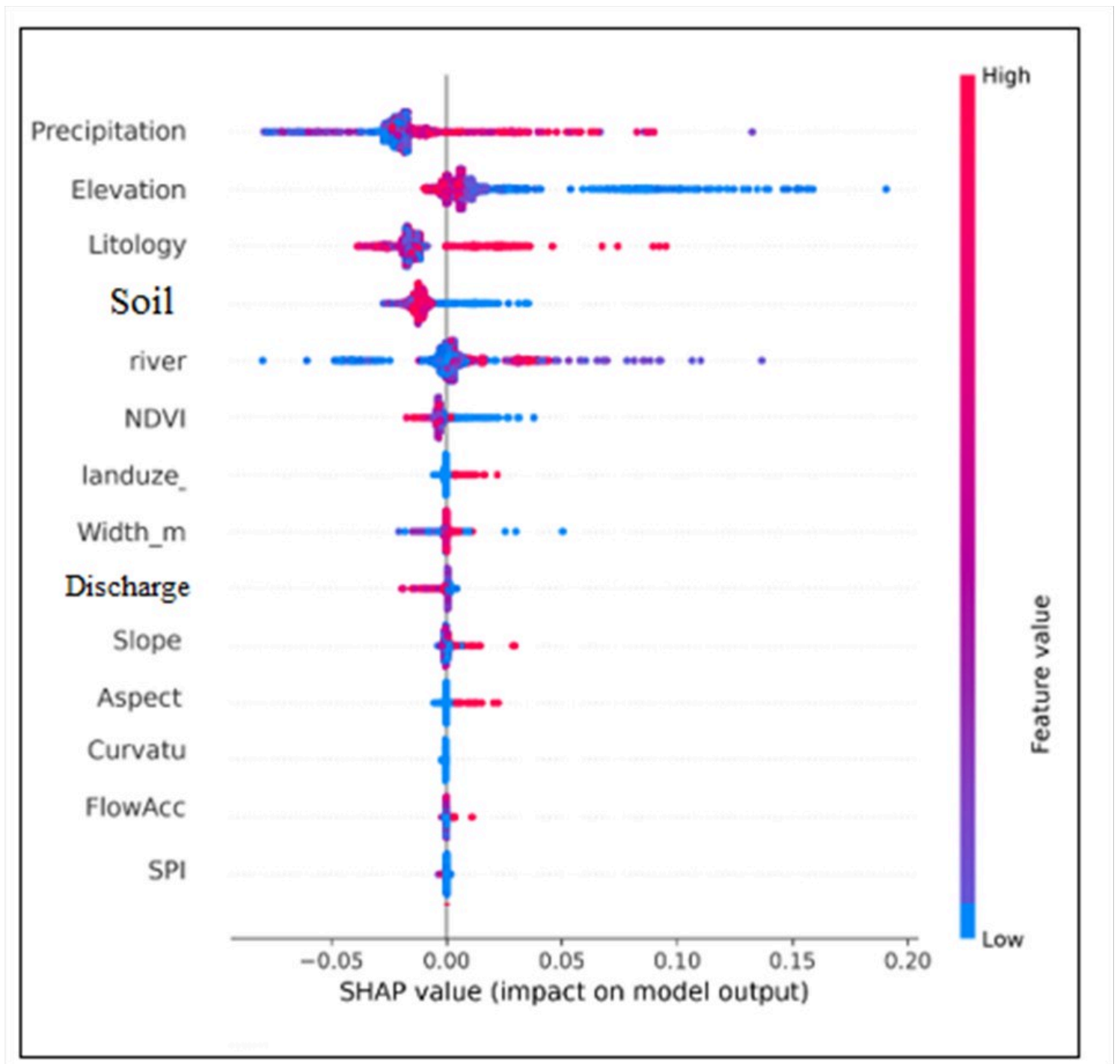


Figure 7

SHAP summary chart

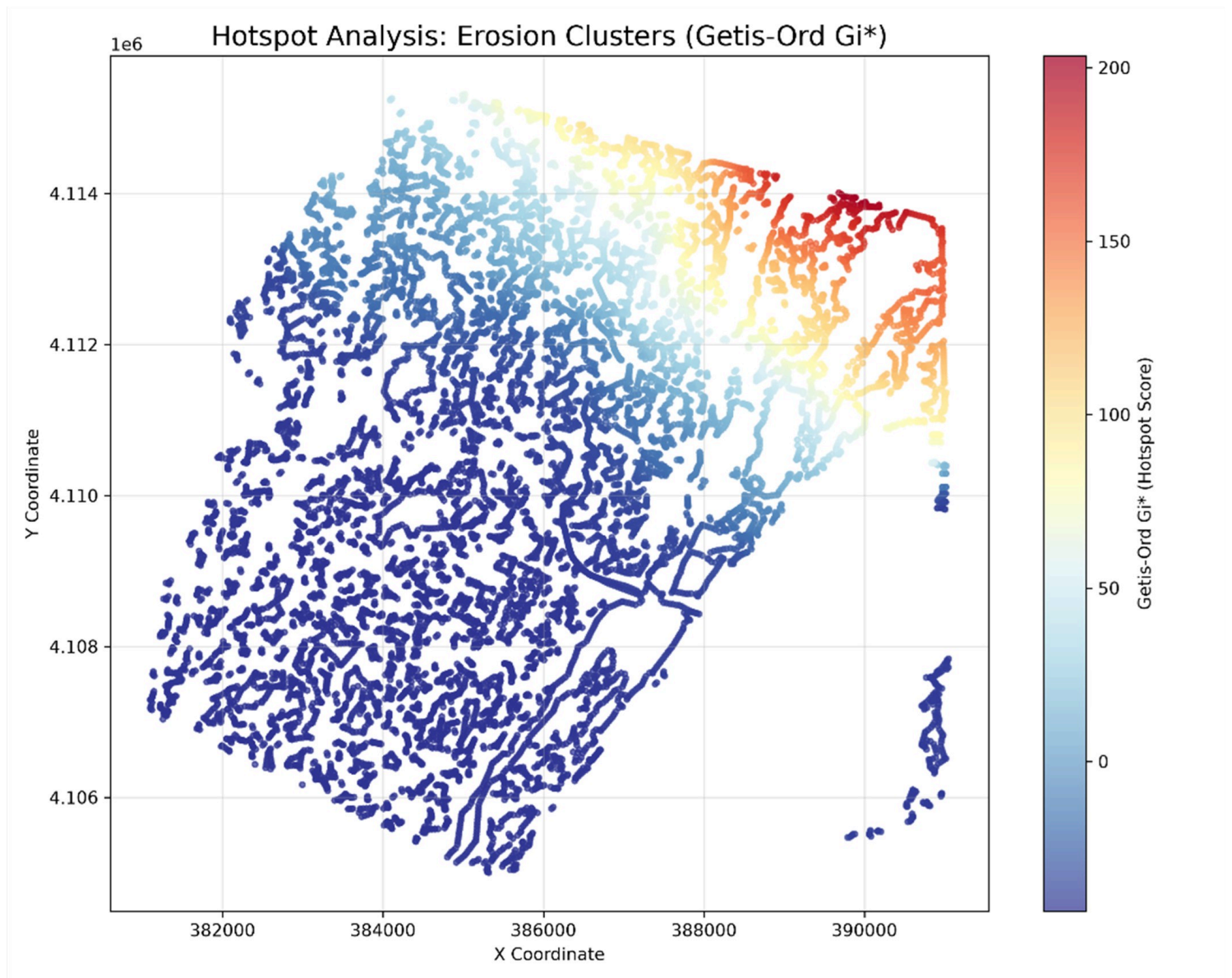


Figure 8

Hotspot map

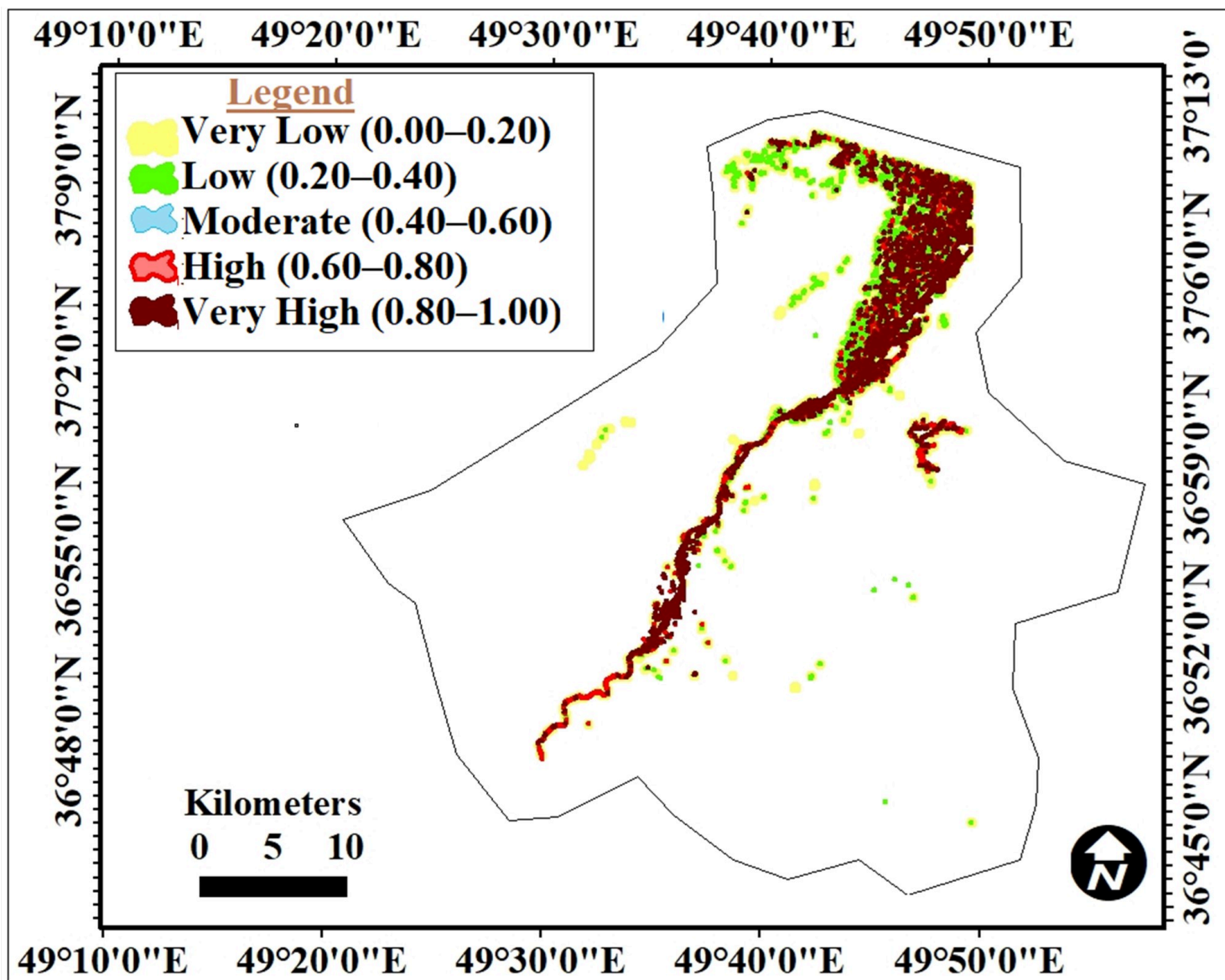


Figure 9

Riverbank erosion susceptibility zoning map