



## OPEN ACCESS

EDITED BY  
Katsuichiro Goda,  
Western University, Canada

REVIEWED BY  
Zarghaam Rizvi,  
GeoAnalysis Engineering GmbH, Germany  
Halil Akinci,  
Artvin Coruh University, Türkiye

\*CORRESPONDENCE  
Xianghang Bu,  
✉ buxh1649@126.com

RECEIVED 15 February 2025  
ACCEPTED 24 March 2025  
PUBLISHED 25 April 2025

## CITATION

Bu X, Fan S, Zhang Z, Zhu K and Ma X (2025)  
Interpretability study of earthquake-induced  
landslide susceptibility combining  
dimensionality reduction and clustering.  
*Front. Earth Sci.* 13:1577165.  
doi: 10.3389/feart.2025.1577165

## COPYRIGHT

© 2025 Bu, Fan, Zhang, Zhu and Ma. This is an  
open-access article distributed under the  
terms of the [Creative Commons Attribution  
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or  
reproduction in other forums is permitted,  
provided the original author(s) and the  
copyright owner(s) are credited and that the  
original publication in this journal is cited, in  
accordance with accepted academic practice.  
No use, distribution or reproduction is  
permitted which does not comply with  
these terms.

# Interpretability study of earthquake-induced landslide susceptibility combining dimensionality reduction and clustering

Xianghang Bu<sup>1,2\*</sup>, Songhai Fan<sup>1,2</sup>, Zongxi Zhang<sup>1,2</sup>, Ke Zhu<sup>1,2</sup> and  
Xiaomin Ma<sup>1,2</sup>

<sup>1</sup>Electric Power Research Institute of State Grid Sichuan Electric Power Company, Chengdu, China,  
<sup>2</sup>Power Internet of Things Key Laboratory of Sichuan Province, Chengdu, China

An earthquake of magnitude Ms5.8 struck Barkam City, Aba Prefecture, Sichuan Province, China, on the morning of 10 June 2022. This was followed by two additional earthquakes of magnitudes Ms6.0 and Ms5.2. The earthquakes triggered significant geological hazards, impacting Barkam City and surrounding areas. Using Random Forest (RF) and Extreme Gradient Boosting (XGBoost) machine learning models, we assessed landslide susceptibility in Barkam City and identified key influencing factors. The study applied the SHAP method to evaluate the importance of various factors, used UMAP for dimensionality reduction, and employed the HDBSCAN clustering algorithm to classify the data, thereby enhancing the interpretability of the models. The results show that XGBoost outperforms RF in terms of accuracy, precision, recall, F1 score, KC, and MCC. The primary factors influencing landslide occurrence are topographic features, seismic activity, and precipitation intensity. This research not only introduces innovative machine learning techniques and interpretability methods for landslide susceptibility analysis but also provides a scientific foundation for emergency response and post-disaster planning related to landslide risks following the earthquake in Barkam City.

## KEYWORDS

landslide susceptibility, random forest, XGBoost, Umap, HDBSCAN

## 1 Introduction

A magnitude Ms5.8 earthquake with a focal depth of 10 km struck Barkam City, Aba Tibetan and Qiang Autonomous Prefecture, Sichuan Province, China, on 10 June 2022 (32.27°N, 101.82°E). The same area was then hit by another earthquake of magnitude Ms6.0 at a depth of 13 km. A third earthquake with a focal depth of 15 km and a magnitude of Ms5.2 happened 2 h later at 32.24°N, 101.85°E (Yue et al., 2024). The Barkam Ms6.0 earthquake swarm is the name given to this sequence of seismic occurrences. A total of 113,950 people were impacted by the earthquakes and the secondary disasters that followed in and around Barkam City. Six people

were injured, and an estimated 2.005 billion yuan was lost directly in economic losses. The incidents resulted in incalculable property damage and human casualties. Therefore, for emergency rescue operations and subsequent resettlement planning, it is essential to quickly and accurately acquire geological disaster distribution after an earthquake (Xuanmei et al., 2022).

Statistical techniques (Lingjing et al., 2023) and physical models (Haijun et al., 2022) are frequently used in the assessment of landslide susceptibility. The former is appropriate for regional studies since it usually creates evaluation models based on probabilistic techniques to determine the likelihood of landslide occurrence. The latter, which is frequently used in case studies, on the other hand, mainly depends on physics and engineering geology knowledge to create conceptual models that mimic the physical processes of landslides. Traditional statistical techniques, which were labor-intensive and produced less-than-ideal results, were primarily employed in early regional research. With the development of artificial intelligence in recent years, machine learning methods such as neural networks (Phuong et al., 2021; Wang et al., 2021), logistic regression (Jiang et al., 2023; Li et al., 2019; Zhang et al., 2022), random forests (Wang et al., 2024; Zhang R. et al., 2024) and support vector machines (SVM) (Huang et al., 2023; Huang and Zhao, 2018) have gained popularity as instruments for determining landslide susceptibility. In order to estimate the likelihood of landslides occurring in particular areas, landslide susceptibility evaluation takes into account the relationships between past landslide disasters and a variety of influencing factors, including geology, topography, soil properties, and human activities. Among the many machine learning techniques, RF and XGBoost have proven to be noticeably better (Xing et al., 2024). By building multiple decision trees and combining their votes, RF successfully lowers the risk of overfitting and shows excellent robustness when working with high-dimensional data (Lin et al., 2024). In the meantime, XGBoost uses ensemble learning strategies to improve model performance. In addition to efficiently capturing intricate nonlinear relationships, it also uses regularization techniques to simplify the model and increase predictive accuracy (Liu and Deng, 2024; Zhu et al., 2024). This makes RF and XGBoost effective tools for assessing landslide susceptibility because they can learn from big datasets and uncover the complex nonlinear relationships affecting the occurrence of landslides while reducing the influence of human subjectivity, leading to more accurate evaluation results.

Machine learning models are frequently viewed as “black boxes” because of their inability to be interpreted, even though they perform exceptionally well and are highly accurate (Faming et al., 2023). To deal with this problem, Lundberg et al. presented a novel approach to the structural interpretation of models and the justification of findings in 2017 with the SHAP method (Lundberg and Lee, 2017). Researchers have started to concentrate on how easily landslide susceptibility assessments can be interpreted (Lei et al., 2024; Xiao et al., 2024; Yang et al., 2024; Yu et al., 2024). The direct interpretation of the internal mechanisms of landslide susceptibility models is complicated by the high dimensionality of the data and the inclusion of multiple evaluation factors. The interpretation of landslide susceptibility research can be greatly improved by using dimensionality reduction techniques, which

can project high-dimensional data into a lower-dimensional space and reveal the intricate nonlinear relationships within the data (Calderon-Guevara et al., 2022). One popular method for reducing dimensionality is Principal Component Analysis (PCA), which preserves the maximum variance information in the data by projecting it onto a small number of principal components using linear transformations. However, the relationships between complex nonlinear features may not be sufficiently revealed by PCA because it assumes linear relationships among features (Sabokbar et al., 2014; Xu et al., 2024). In this regard, nonlinear dimensionality reduction techniques such as Uniform Manifold Approximation and Projection (UMAP) and t-distributed Stochastic Neighbor Embedding (t-SNE) have attracted growing interest (Kobak and Linderman, 2021; Marx, 2024). High-dimensional data can be efficiently mapped into a lower-dimensional space using t-SNE, which is especially useful for displaying the local structure of high-dimensional data by enabling similar data points to group together in the lower-dimensional space. However, the results of t-SNE can be greatly impacted by the choice of parameters, and it does not perform well in maintaining global structure (Han et al., 2022). The interpretability studies of landslide susceptibility, on the other hand, show that UMAP is more capable of preserving both local and global structures in the lower-dimensional space (Weaver et al., 2022).

The implementation of clustering methods in landslide susceptibility research greatly improves the comprehensibility of models (Zhao X. et al., 2024). Specifically, HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), a sophisticated clustering method, excels at autonomously identifying areas of varying data density (Stewart and Al-Khassawneh, 2022). Unlike conventional clustering techniques like K-means, HDBSCAN is adept at handling data with diverse shapes and densities, and it can dynamically discover potential clustering structures (Wang et al., 2023). This attribute makes it a valuable tool for landslide susceptibility studies, as geological data frequently display intricate and irregular spatial patterns. HDBSCAN can effectively segment these data into numerous subsets with similar traits, enabling researchers to conduct a more in-depth analysis of the roles and impacts of various evaluation elements within each cluster. This detailed analysis helps in comprehending the significance of various factors in landslide occurrence and uncovers potential influencing mechanisms, thereby enhancing the model's interpretability.

This study takes into account the region's complex natural and geographical conditions to assess landslide susceptibility. Thirteen key indicators are selected for the landslide susceptibility evaluation: elevation, slope, curvature, precipitation, lithology, land use type, distance to roads, distance to water systems, distance to fault lines, NDVI (Normalized Difference Vegetation Index), TWI (Topographic Wetness Index), SPI (Stream Power Index), and PGA (Peak Ground Acceleration). The landslide susceptibility of the Barkam region is assessed using Random Forest and XGBoost machine learning models. By integrating SHAP, HDBSCAN, and UMAP techniques, this study enhances the understanding of how various evaluation factors influence landslide occurrences. This integration not only improves the interpretability of the landslide susceptibility assessment but also

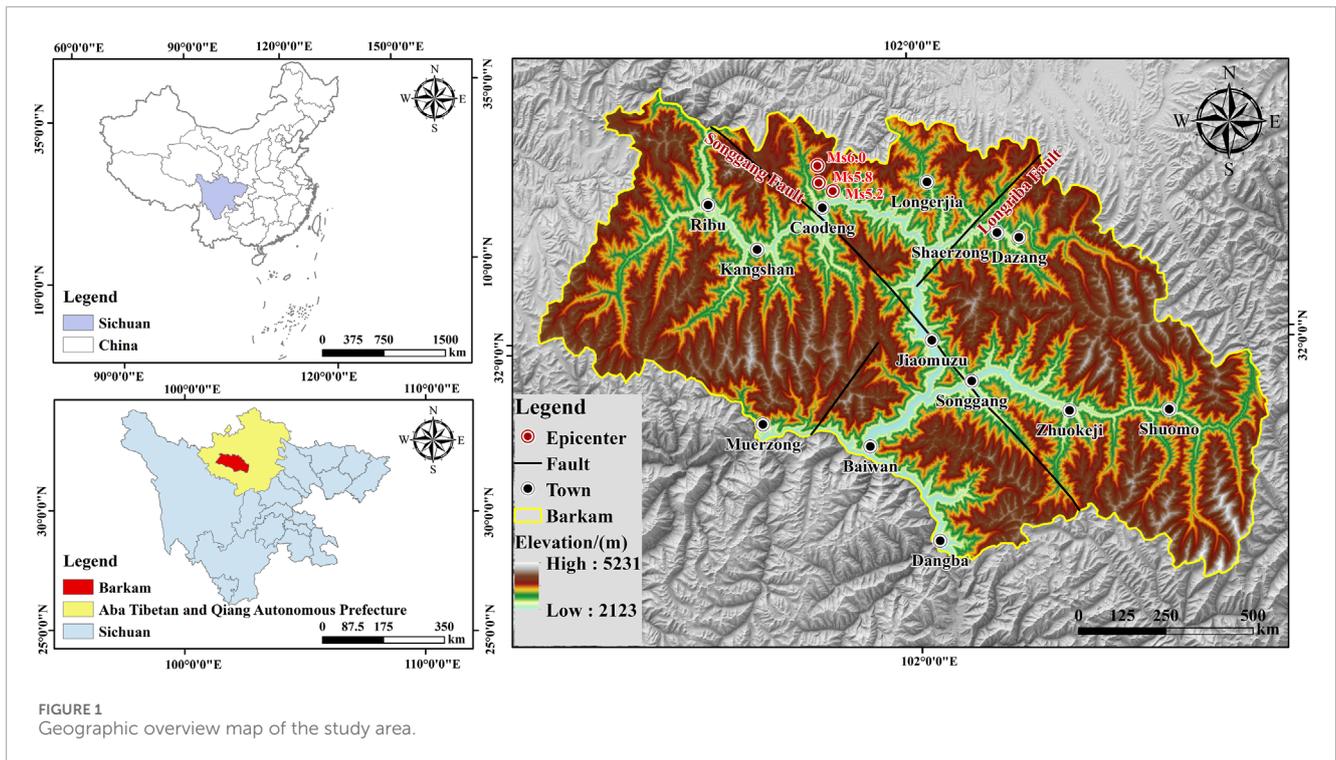


FIGURE 1  
Geographic overview map of the study area.

strengthens the models' accuracy and stability, making them a reliable tool for emergency response and post-earthquake planning.

## 2 Study area

Barkam City, located at the intersection of the Qinghai-Tibet Plateau and the Sichuan Basin, serves as the capital of the Aba Tibetan and Qiang Autonomous Prefecture in Sichuan Province, China. The city covers an area of 6,622.88 km<sup>2</sup> and is situated between the geographic coordinates of 101°20' to 102°50' E and 31°35' to 32°20' N. The complex terrain of Barkam, which slopes down from the northeast to the southwest, is characteristic of a plateau mountain landform, featuring a mix of hills and mountains. The study area belongs to alpine canyon geomorphology and the overall topography is steep with a maximum slope of 80.59° and an average slope of 28.98°. The region's unique geographical setting, combining low latitude and high altitude, sees elevations ranging from 2,123 m at its lowest point to 5,231 m at its highest. The lithologic of the study area is dominated by sandstone, granite, kyanite, basalt, gray rock, etc. In addition, many sand pebbles, gravel soil and other Q4 loose soil layers distributed along the banks of the river. The Barkam Ms6.0 earthquake swarm occurred at the junction of the Songgang and southern Longriba faults. Due to its well-developed secondary faults and complex tectonic structure, the area experiences frequent moderate to small earthquakes. As shown in Figure 1, the Songgang fault zone consists of several secondary faults, running in a NW-SE direction (Dong et al., 2024).

## 3 Data and research methodology

### 3.1 Data

Previous studies have examined soil properties under various thermal and mechanical conditions, including their evolution under cyclic loading (Ahmad et al., 2019; Ahmad et al., 2025; Ahmad et al., 2021). These considerations are significant when evaluating soil behavior in landslide susceptibility models. The distribution of landslide sites in Barkam city and the zoning map of peak seismic acceleration for the Barkam Ms6.0 magnitude earthquake swarm were provided by the Geological Survey of Sichuan Province. In accordance with the conditions that cause landslides, 13 evaluation factors were chosen from five aspects of topography and geomorphology, precipitation, geology, vegetation cover, and human activities. (1) Topographic and Geomorphologic data, based on the DEM data with a spatial resolution of 12.5 m in the study area, were extracted by Arcgis10.8 to obtain the elevation, slope, curvature, and water system; (2) Precipitation data, the average annual precipitation vector data of the study area were obtained from the National Meteorological Science Data Center; (3) Geology: Fracture zone data and stratigraphic rock data of the study area were obtained from the Resource and Environment Science and Data Center of the Chinese Academy of Sciences (<http://www.resdc.cn/>), and the distance to the fracture zone was obtained by converting coordinates; (4) Vegetation cover data, the Landsat8 near-infrared and far-infrared bands with a resolution of 30 m in June 2022 were selected to obtain the NDVI; the vector data of land use types in the study area were obtained by cropping the global 30-meter surface cover data (<http://www.globeland30.org/home/background.aspx>); (5) Data on human activity: the 2021

TABLE 1 Statistical tables of data sources.

Category	Feature factor	Indicator meaning	Data source
Topography	Slope/(°)	Indicates the degree of slope and surface inclination	DEM (12.5 m resolution), Geospatial Data Cloud
	Curvature	Indicates the local variation in surface undulation	
	Topographic Relief/(m)	Indicates the variation in elevation difference	
Hydrology	Distance to Rivers/(m)	Distance to rivers	Based on 1:250,000 basic geographic information data
	Topographic Wetness Index	Indicates soil moisture content and humidity conditions	DEM (12.5 m resolution), Geospatial Data Cloud
	Stream Power Index	Indicates the erosion effect of water flow on slopes	
Seismic	Peak Ground Acceleration/(gal)	Indicates the impact of the maximum amplitude of seismic acceleration time history on slopes	Based on data released by the China Earthquake Networks Center
Human Activities	Distance to Roads/(m)	Distance to roads	Based on 1:250,000 basic geographic information data
Geological Structure	Distance to Faults/(km)	Distance to faults	Based on data released by the China Geological Survey Data Service Platform
	Rock Hardness	Indicates the hardness of different rock types	Based on geological maps and rock hardness level qualitative classification tables downloaded from the China Geological Survey Data Service Platform
Environmental Geological Characteristics	Land Type	Indicates different land types affecting slope development	Based on Landsat8 remote sensing image data, Geospatial Data Cloud
	Normalized Difference Vegetation Index (NDVI)	Indicates vegetation growth and coverage within a range	

Gaofen-2 satellite image was used to extract information about the study area's roads and as well as the distance from each road. Table 1 and Figure 2 displays the distribution of landslide hazard points and evaluation factors (6).

Landslide occurrences in Malcolm City were interpreted using high-resolution Landsat eight satellite imagery from Google Maps. The interpretation identified 1,142 landslides (Figure 2) covering a total area of approximately 96.84 km<sup>2</sup>, representing 1.7% of the total image area. These landslides were categorized by scale: 891 small-scale, 242 medium-scale, and nine large-scale, with no very-large landslides identified. The majority were classified as soil landslides, primarily composed of materials such as silty clay, gravelly soil, and stony soil. Geographically, they were predominantly distributed across five townships: Caodeng (515 landslides), Ribu (160 landslides), Long'erjia (191 landslides), Jiamuzu (103 landslides), and Kangshan (104 landslides).

## 3.2 Methodology

### 3.2.1 Random forest

Random Forest (RF) is an ensemble learning algorithm that builds several decision trees and aggregates their predictions to improve overall predictive accuracy. Training sets are created by randomly selecting several subsets from the original dataset, and a decision tree is constructed for each training set using a randomly selected subset of features. Each decision tree is constructed using a random feature selection process, which reduces overfitting. The training set for each decision tree is used until a stopping criterion is satisfied. When making predictions about new data, the Random Forest algorithm aggregates the predictions made by each decision tree using the Bagging ensemble approach. It then uses voting or averaging to determine the final prediction value for regression or classification (Zhao Z. et al., 2024). Numerical modeling techniques, such as lattice element modeling, have been

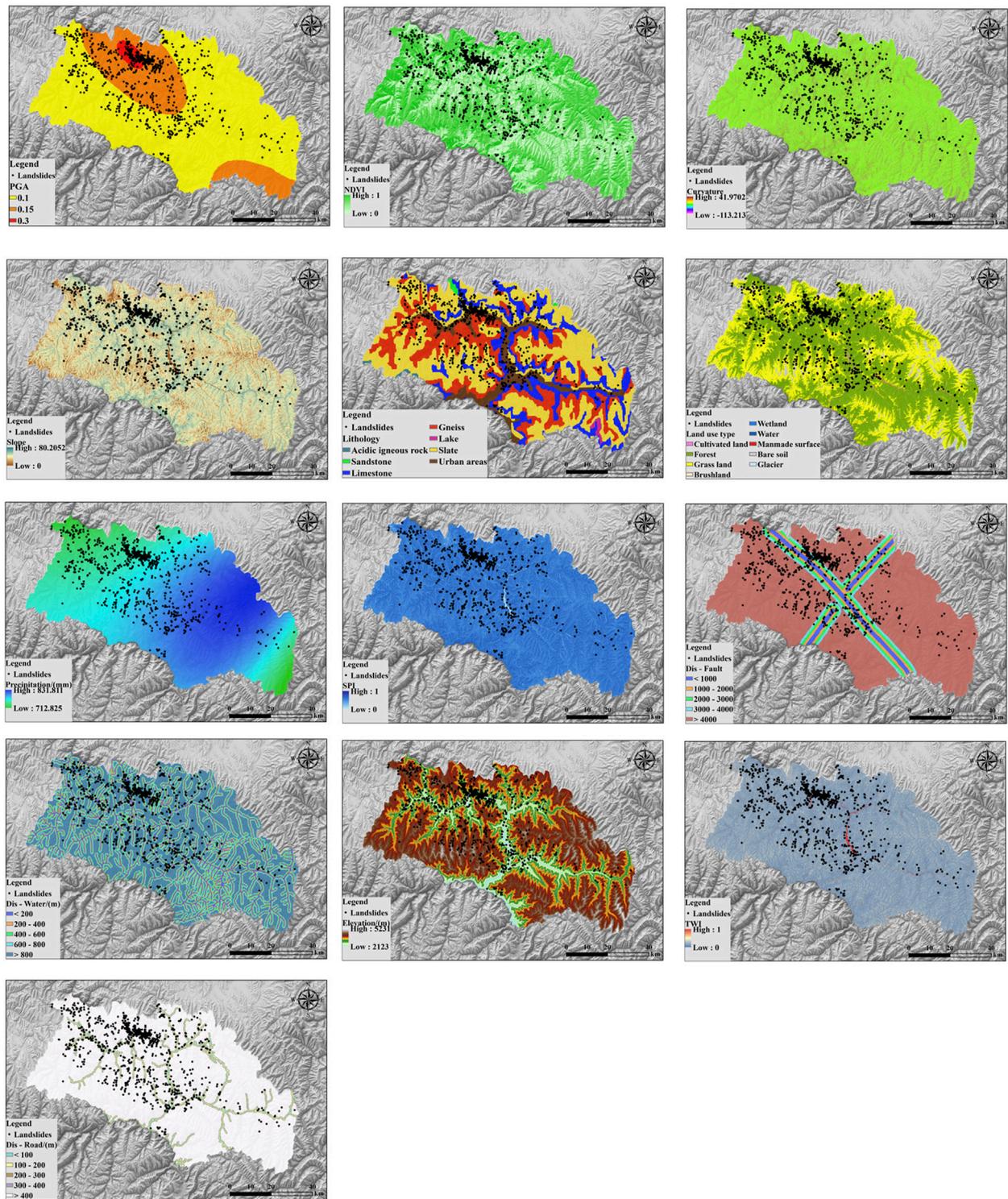


FIGURE 2 Evaluation factors and distribution of landslide hazard sites.

widely used for geomechanically failure simulations (Rizvi et al., 2020), offering an alternative perspective to machine learning-based susceptibility models.

### 3.2.2 XGBoost

Based on Gradient Boosting Decision Trees (GBDT), Extreme Gradient Boosting (XGBoost) is an optimized algorithm. The

serial algorithms XGBoost and conventional GBDT both use the Boosting ensemble approach, in which each decision tree is trained using the residuals between the values that were observed and the predictions of the preceding tree. Over several iterations, the algorithm fits the residuals and gets closer to the real values. Training ends when a predefined value or number of iterations is reached, and the weighted summation of all decision tree predictions yields the final prediction for a sample. The main distinction is in how the loss function is calculated: XGBoost uses second-order Taylor expansion to make computations easier and adds regularization terms, like L1 and L2, to the objective function to mitigate overfitting and control model complexity. (Zhang et al., 2023). Similar optimization techniques have been employed in computational fluid dynamics models to improve efficiency and accuracy in large-scale simulations (Haroon et al., 2017).

### 3.2.3 UMAP

UMAP is a nonlinear dimensionality reduction algorithm that employs similarity theory to map high-dimensional data to a lower dimension while preserving the global and local topological structures of the data as much as possible. This characteristic allows it to perform exceptionally well across various datasets (McInnes and Healy, 2018). The landslide susceptibility evaluation indicator sample dataset exhibits nonlinear high-dimensional features, and the computational steps for applying UMAP for dimensionality reduction are as follows:

- (1) Let the input landslide susceptibility evaluation indicator sample dataset be denoted as  $X = \{x_1, \dots, x_i, \dots, x_n\}$ , for each  $x_i$ , we utilize the nearest neighbor descent method to obtain  $\{x_{i1}, \dots, x_{ik}\}$ , where  $k$  is a hyperparameter, define  $x_i$  as follows (Equations 1, 2):

$$\rho_i = \min \{d(x_i, x_{ij}) \mid 1 \leq j \leq k, d(x_i, x_{ij}) > 0\} \tag{1}$$

$$\sum_{j=1}^k \exp\left(\frac{-\max(0, d(x_i, x_j) - \rho_i)}{\sigma_i}\right) = \log_2 k \tag{2}$$

Where  $\rho_i$  is chosen to connect  $x_i$  with at least one neighbor with a weight of 1;  $\sigma_i$  is a length parameter;  $d()$  is a metric defined as  $d: X \times X \rightarrow R \geq 0^*$ .

- (2) Next, we define a directed weighted graph  $\bar{G} = (V, E, w)$ . Utilizing its symmetry, we define an undirected weighted graph  $G$ , where the vertex set  $V$  corresponds to the set  $X$ , leading to the directed edge set  $E = \{(x_i, x_{ij}) \mid 1 \leq j \leq k, 1 \leq i \leq N\}$ . The weight function is defined as Equations 3, 4:

$$w(x_i, x_{ij}) = \exp\left(\frac{-\max(0, d(x_i, x_{ij}) - \rho_i)}{\sigma_i}\right) \tag{3}$$

Let  $A$  be the adjacency matrix of  $\bar{G}$ . The adjacency matrix  $B$  of the undirected weighted graph  $G$  is then given by:

$$B = A + A^T - A \otimes A^T \tag{4}$$

Where  $\otimes$  denotes the Hadamard product.

- (3) UMAP applies attractive and repulsive forces along the edges and vertices, evolving into an equivalent weighted graph  $H$ , which includes the point set  $\{y\}$ . The attraction and repulsion at points  $y_i$  and  $y_j$  are defined as follows (Equations 5, 6):

$$GR = \frac{-2ab \|y_i - y_j\|_2^{2(b-1)}}{1 + \|y_i - y_j\|_2^2} \omega(x_i, x_j)(y_i - y_j) \tag{5}$$

$$RE = \frac{2b(1 - \omega(x_i, x_j))(y_i - y_j)}{(\epsilon + \|y_i - y_j\|_2^2)(1 + a \|y_i - y_j\|_2^{2b})} \tag{6}$$

Where  $GR$  represents the attraction force;  $RE$  denotes the repulsion force;  $\epsilon$  is a small number added to prevent division by zero;  $a$  and  $b$  are hyperparameters.

### 3.2.4 HDBSCAN

HDBSCAN is an algorithm proposed by Campello et al. that combines density-based clustering and hierarchical clustering. The algorithm introduces a measure of mutual reachability distance to construct a hierarchical structure for different clusters, enabling clustering of groups with varying densities (Campello et al., 2013). HDBSCAN can evaluate the membership degree for each sample, with a range of  $[0, 1]$ . A membership degree of 0 indicates that the sample point is a noise point and does not belong to any cluster; a membership degree of one indicates that the sample point is a core point, and the attributes of the cluster core point can represent the typical characteristics of that cluster. The steps of the algorithm are as follows:

- (1) Spatial Transformation of Data Points: The core distance  $c_k(x) = d(x, N^k(x))$  indicates the distance from the data point  $x$  to its  $k$ -th nearest neighbor. The mutual reachability distance between two points is calculated as Equation 7:

$$d_k(a, b) = \max \{c_k(a), c_k(b), d(a, b)\} \tag{7}$$

Where  $d(a, b)$  refers to the straight-line distance between points  $a$  and  $b$ .

- (2) Construction of the Minimum Spanning Tree: The HDBSCAN algorithm internally uses Prim's algorithm to treat the sample points as original points, using the mutual reachability distances to other points as weights to construct the minimum spanning tree.
- (3) Establishment of Cluster Hierarchical Structure: The algorithm traverses and reorders the edges of the minimum spanning tree using mutual reachability distances as weights, categorizing each edge into a new cluster.
- (4) Compression of the Clustering Tree: When compressing and segmenting the coarse hierarchical structure, the algorithm compares the number of samples in the newly segmented clusters with the minimum cluster size, removing the smaller one.
- (5) Extraction of Clusters: The algorithm uses  $\lambda = 1/d$  as a standard to measure the clustering results and introduces three quantities:  $\lambda_{\text{birth}}$ ,  $\lambda_{\text{death}}$ ,  $\lambda_p$ , which represent the formation of a cluster, the splitting of a cluster into two

TABLE 2 Model accuracy evaluation indicators.

Index	Formulas	Description
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$	Calculating the percentage of samples that were accurately predicted
Precision	$\frac{TP}{TP+FP}$	Calculating the TP sample percentage in each predicted positive sample
Recall	$\frac{TP}{TP+FN}$	Calculating the TP sample's percentage in each true positive sample
F1 Score	$\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$	Representing the accuracy and recall harmonic mean, with a range of values from -1 to 1
KC	$P_e = \frac{\frac{P_0 - P_1}{1 - P_1} \frac{TP+TN}{(TP+FN)(TP+FP)+(TN+FN)(FP+TN)}}{(TP+FN+FP+TN)^2}$	Checking consistency and measuring classification precision
MCC	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$	Describing the correlation coefficient between the actual classification and the predicted classification, with a value range of -1 to 1

sub-clusters, and the value when separating from the parent cluster, respectively. The stability of the clusters is evaluated using the stability measure  $S_{cluster}$ . The calculation of  $S_{cluster}$  is given by Equation 8:

$$S_{cluster} = \sum_{p \in C_{cluster}} (\lambda_p - \lambda_{death}) \tag{8}$$

Where  $C_{cluster}$  is the set of clusters.

### 3.2.5 SHAP

The SHAP algorithm utilizes the concept of Shapley values from game theory to precisely calculate the contribution of each feature to the model's output. It provides an intuitive representation of each feature's weight and influence in the model's predictions (Lundberg and Lee, 2017).

For a feature  $i$  in the feature set  $S$ , the Shapley value is computed using the following formula (Equation 9):

$$\Phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (\nu(S \cup \{i\}) - \nu(S)) \tag{9}$$

Where  $N$  represents the set of all features;  $S$  is any subset of features that does not include feature  $i$ ;  $|S|$  denotes the number of features in set  $S$ ;  $\nu(S)$  is the contribution of the feature set  $S$  to the model's prediction output; and  $\nu(S \cup \{i\})$  is the contribution of the feature set  $S \cup \{i\}$  to the model's prediction output.

### 3.2.6 Silhouette coefficient

Based on the intrinsic structural features of the clusters, the silhouette coefficient is frequently used in clustering to assess the quality of the clustering results. Better clustering performance is indicated by higher values, which range from -1 to 1. (Lianjiang et al., 2010). The following is the definition of the particular formula (Equation 10):

$$s(i) = \frac{b(i) - a(i)}{\max a(i), b(i)} \tag{10}$$

Where  $a(i)$  denotes the average distance from the  $i$ -th object to all other objects in the same cluster, and  $b(i)$  denotes the average distance from the  $i$ -th object to all objects in the nearest different cluster. The value  $s(i) \in [-1, 1]$ , and the closer this value is to 1, the more reasonable the clustering classification.

### 3.2.7 Model accuracy evaluation

#### 3.2.7.1 Accuracy, precision, recall, F1 score, KC and MCC

An essential component of landslide susceptibility assessment is model validation and performance evaluation. Typically, the performance of binary classification models is evaluated using a confusion matrix. This matrix consists of four key parameters: True Positive (TP), which represents the instances where the model correctly predicts a landslide; False Negative (FN), representing instances where the model predicts a non-landslide but the actual event is a landslide; False Positive (FP), indicating instances where the model predicts a landslide but the actual event is a non-landslide; and True Negative (TN), representing instances where the model correctly predicts a non-landslide (Chen S. et al., 2024). Based on these parameters, the performance of each model is evaluated using six statistical indicators: accuracy, precision, recall, F1 score, Kappa coefficient (KC), and Matthews correlation coefficient (MCC). Table 2 presents the description of each indicator.

#### 3.2.7.2 ROC values and AUC curves

The True Positive Rate (TPR) is shown on the vertical axis of the Receiver Operating Characteristic (ROC) curve, and the False Positive Rate (FPR) is shown on the horizontal axis. It displays how well the model performs at various classification thresholds (Zhang X. et al., 2024). The False Positive Rate refers to the percentage of negative samples incorrectly predicted as positive, while the True Positive Rate represents the percentage of positive samples correctly predicted as positive. The model's performance improves as the ROC curve approaches the upper-left corner.

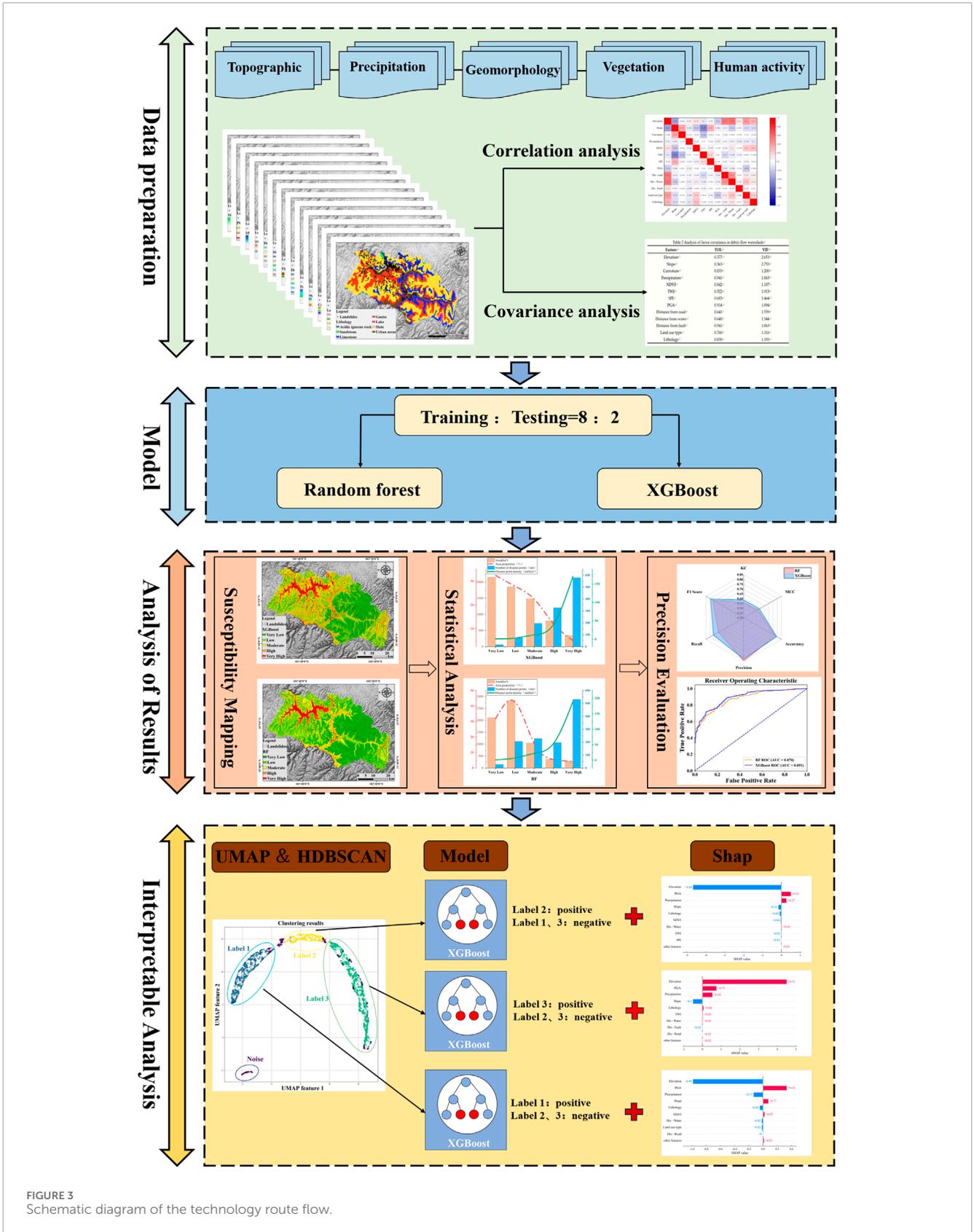


FIGURE 3 Schematic diagram of the technology route flow.

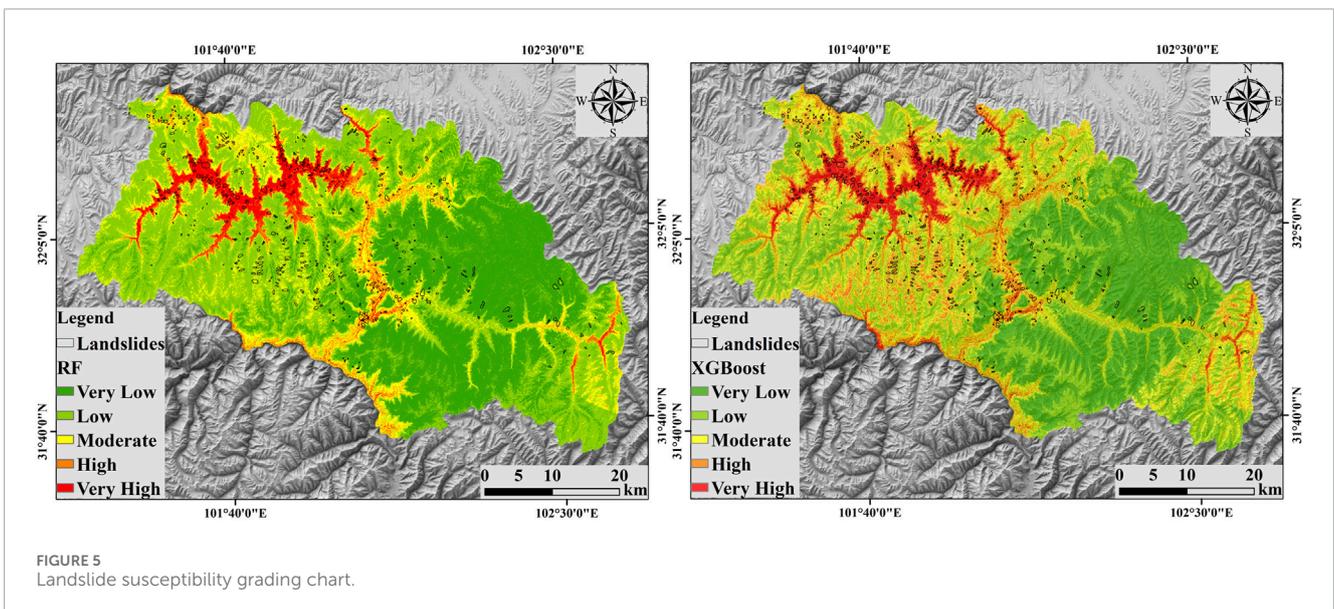
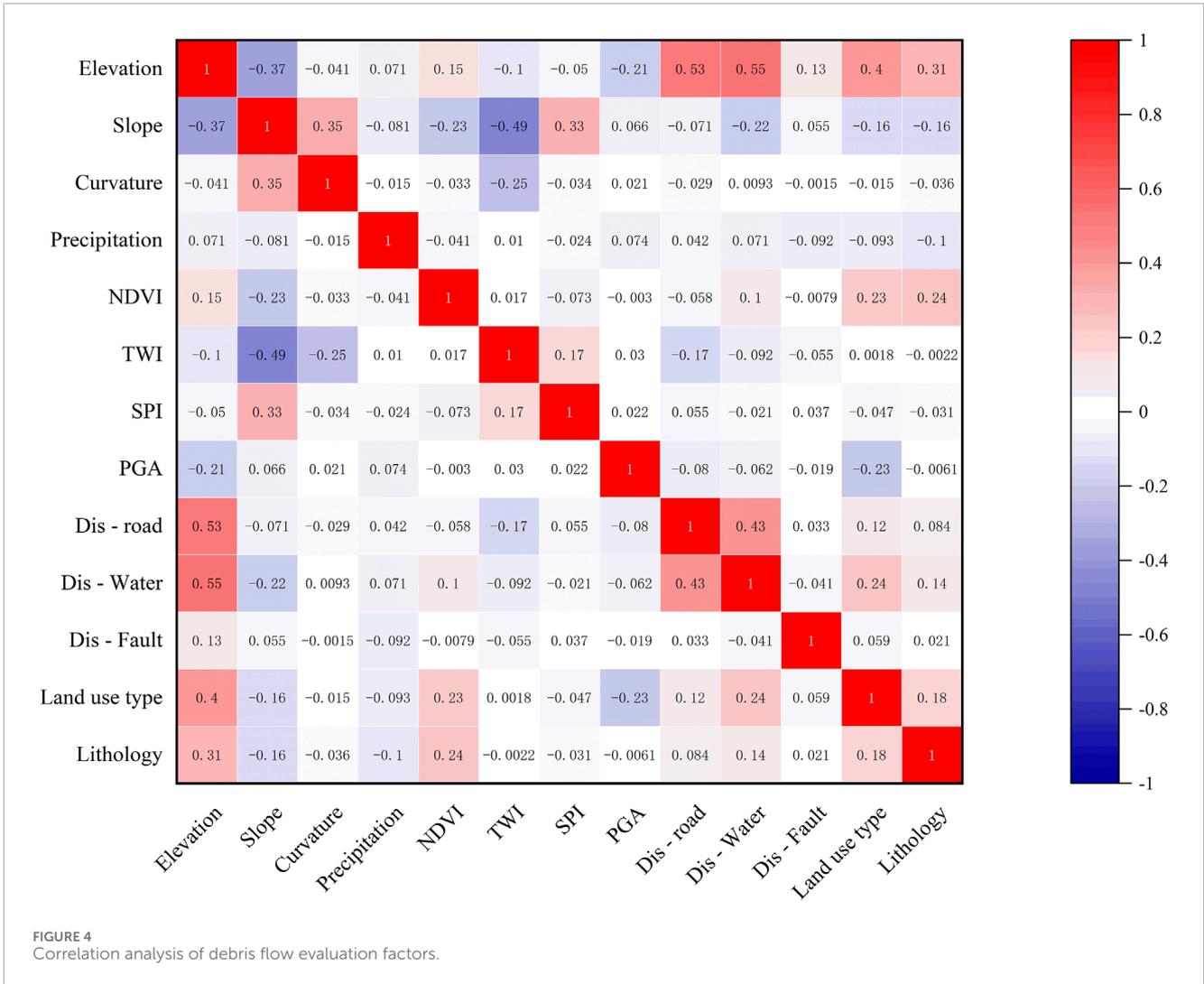


TABLE 3 Statistics of landslide susceptibility class and landslide density for different models.

Model	Classification	Area/km <sup>2</sup>	Area proportion/%	Number of landslide	Proportion of landslide/%	Landslide density/(ind/km <sup>2</sup> )
XGBoost	Very low	2,148.05	32.43	17	1.43	0.79
	Low	1848.78	27.92	73	6.13	3.95
	Moderate	1,486.00	22.44	195	16.39	13.12
	High	800.64	12.09	326	27.39	40.72
	Very high	339.40	5.12	579	48.66	170.59
RF	Very low	2,107.25	31.82	27	2.27	1.28
	Low	2,828.27	42.70	206	17.31	7.28
	Moderate	1,043.52	15.76	228	19.16	21.85
	High	362.12	5.47	197	16.55	54.40
	Very high	281.72	4.25	532	44.71	188.84

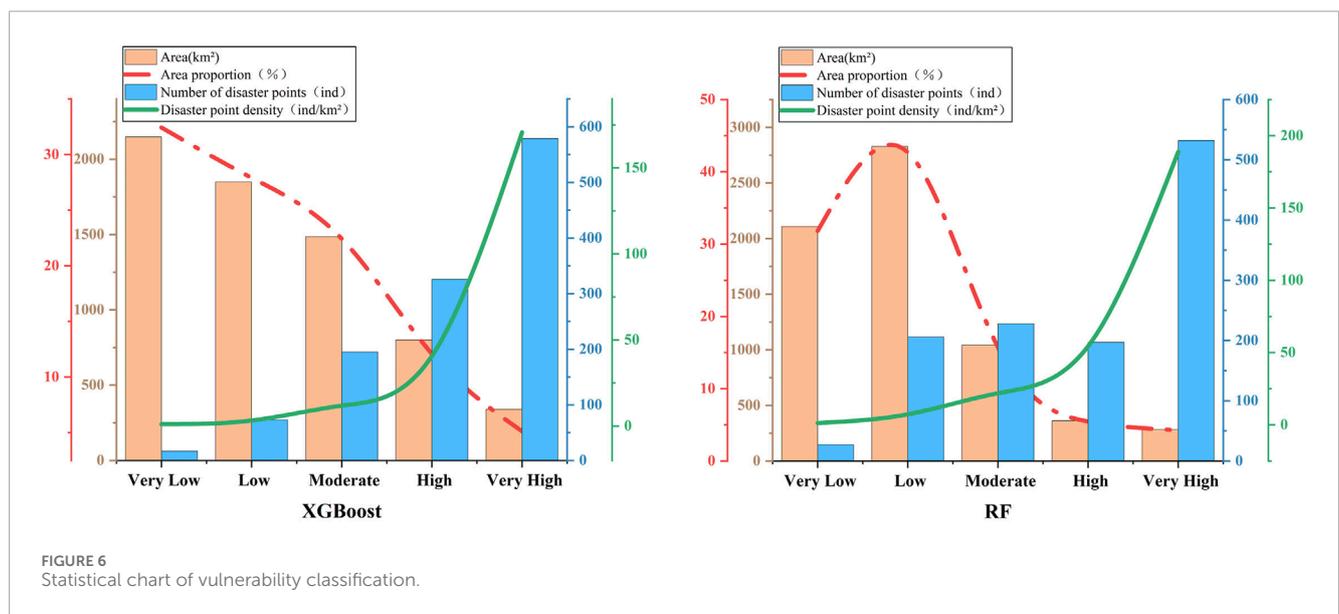


TABLE 4 Statistics of model performance metrics.

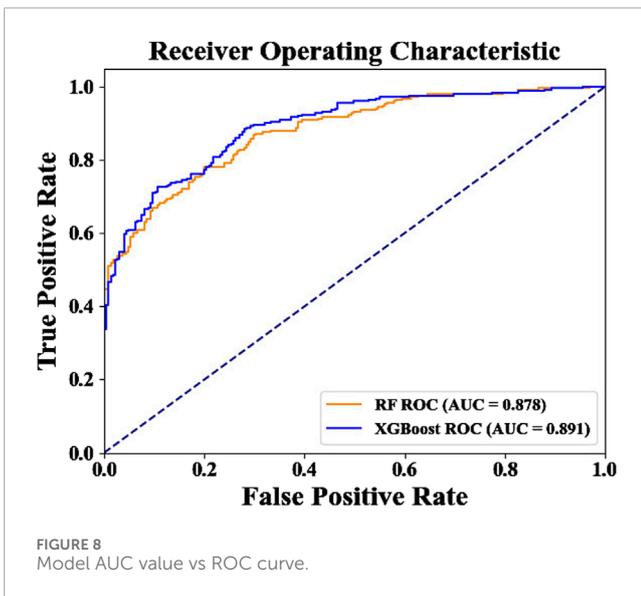
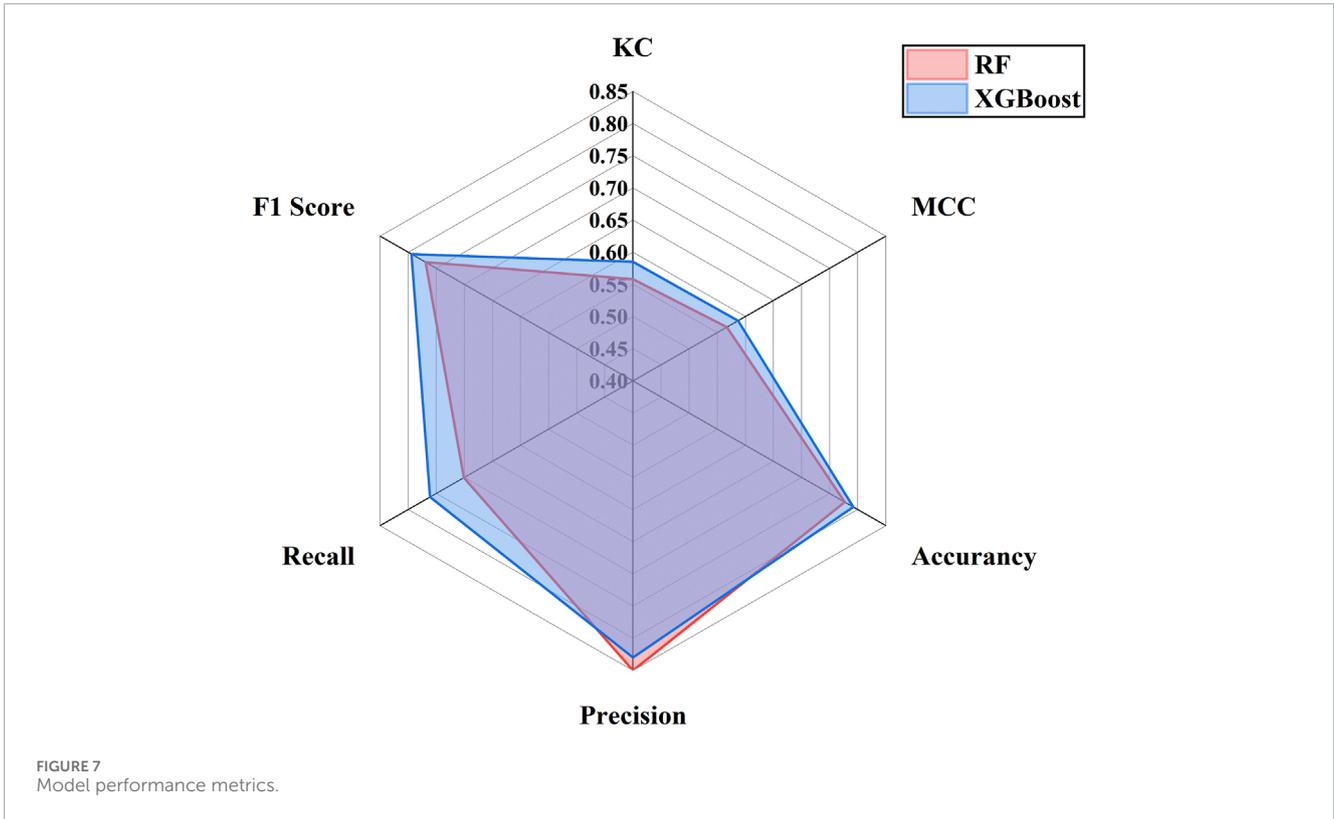
Model	RF	XGBoost
Accuracy	0.777	0.792
Precision	0.85	0.83
Recall	0.701	0.761
F1 Score	0.769	0.794
KC	0.558	0.585
MCC	0.567	0.587

### 3.2.8 Research process

To better understand landslide susceptibility and its influencing factors, this study evaluates landslide susceptibility in the Barkam region using RF and XGBoost models, along with UMAP for dimensionality reduction and HDBSCAN clustering. The research process includes data preparation, model construction, result analysis, and interpretability analysis.

#### 3.2.8.1 Data preparation

Soil water retention and thermal capacity play a critical role in landslide behavior, particularly in mixed-grain soils with variable gravel content (Beck-Broichsitter et al., 2023), affecting their response to prolonged saturation and seismic events. Thirteen



assessment factors from five categories (vegetation cover, geology, precipitation, terrain, and human activities) were selected to reflect the conditions. Multicollinearity and correlation analyses were performed on the data. Using high-resolution imagery and field data, 1,190 landslide events were identified. Non-landslide regions were determined via GIS, and 1,190 non-landslide points were randomly generated to balance the dataset. The final dataset used 80% for training and 20% for testing.

### 3.2.8.2 Model construction

No normalization or standardization was applied to the features, as decision-tree-based ensemble models (RF and XGBoost) are insensitive to feature scaling. The 13 factors were extracted in ArcGIS, and susceptibility models were built using the Scikit-learn package.

### 3.2.8.3 Result analysis

GIS technology was used to map landslide susceptibility levels, and various evaluation metrics (accuracy, precision, recall, F1 score, MC, KCC) were computed on the test set. The AUC value was also calculated for a comprehensive performance assessment.

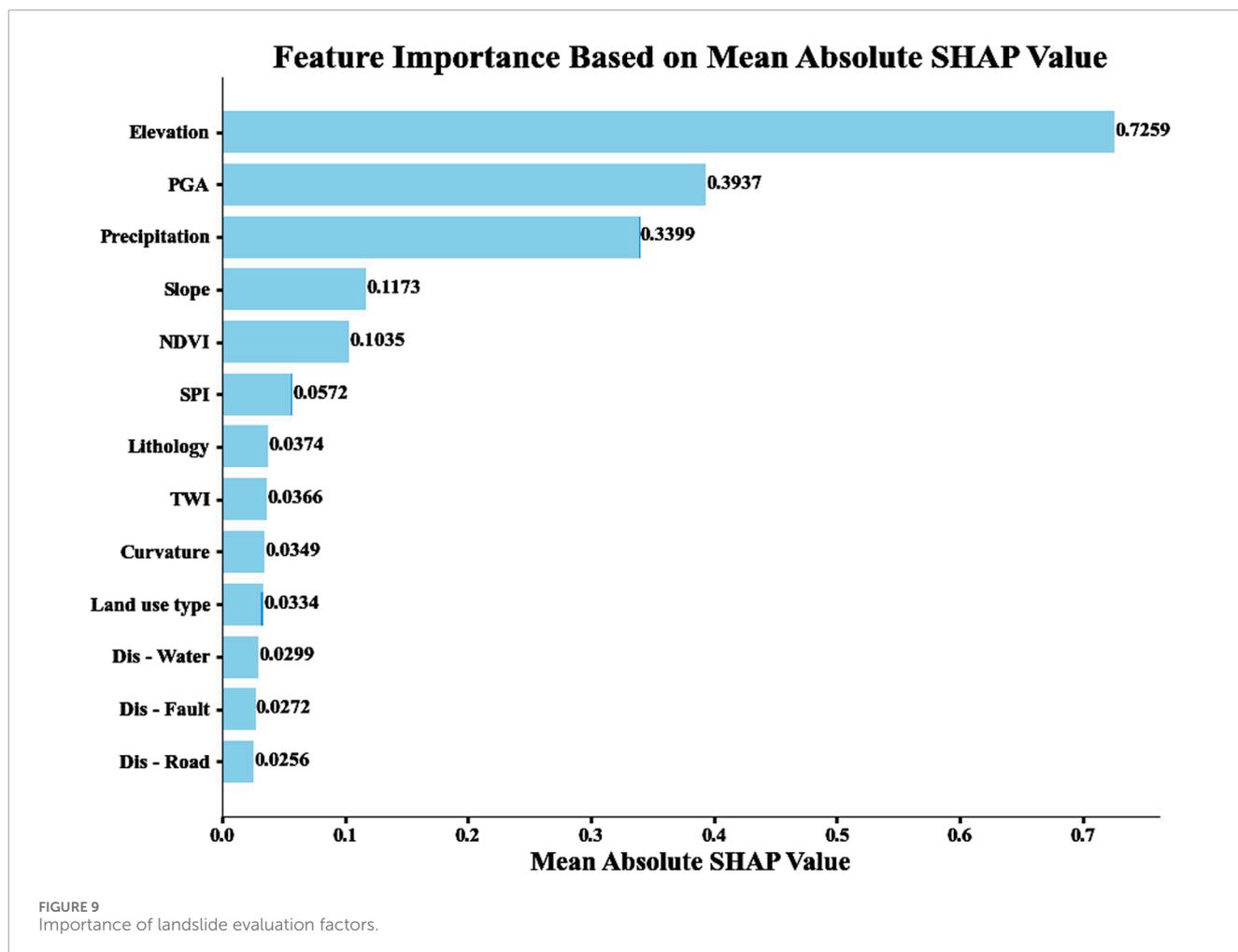
### 3.2.8.4 Interpretability analysis

SHAP analysis was integrated with the XGBoost model to assess factor importance, UMAP was used for dimensionality reduction, and HDBSCAN clustering was applied. One-to-many training and prediction techniques were then employed. Figure 3 illustrates the technical workflow of the study.

## 4 Landslide susceptibility analysis

### 4.1 Correlation test

A correlation analysis was conducted on the 13 factors that were initially chosen in order to identify the most predictive evaluation factors and enhance the prediction accuracy of the model. A correlation matrix for the 13 influencing factors was created using the Origin plotting software's Correlation Plot plugin, and Figure 4



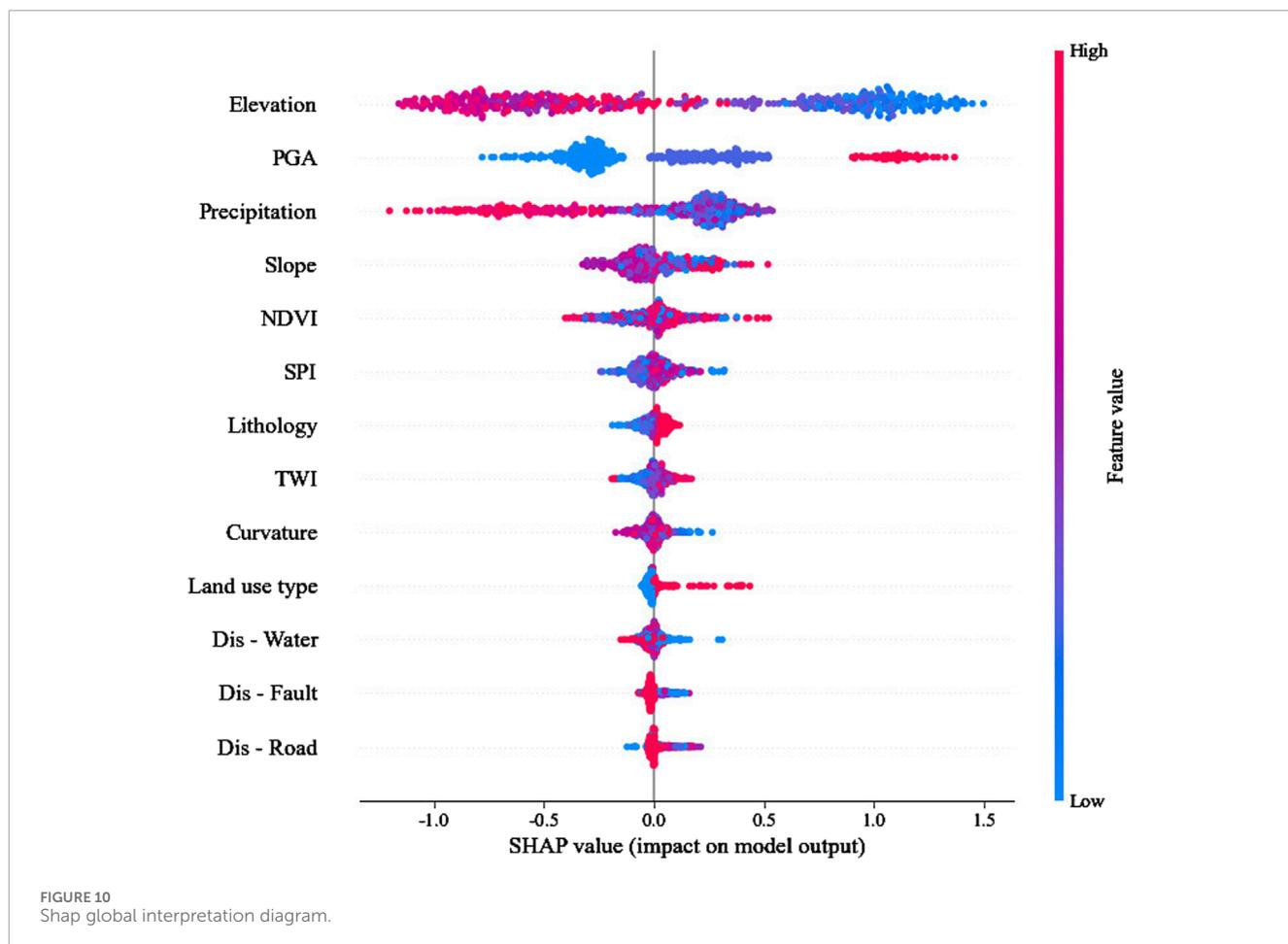
shows the outcomes. Red indicates positive correlations, and blue indicates negative correlations. The size of the plot's color intensity and the correlation coefficient's size are directly correlated. All of the evaluation factors have correlation coefficients below 0.55, which indicates comparatively weak correlations, as can be seen from the plot. This implies that there is little interaction between the factors and that the evaluation factors chosen make sense for the model.

## 4.2 Landslide susceptibility mapping and analysis

The landslide susceptibility index for every evaluation unit in the study area was determined following the stabilization of the XGBoost and RF-based landslide susceptibility models. The susceptibility index was separated into five levels using ArcGIS 10.8's Natural Breaks classification method: very low susceptibility, low susceptibility, moderate susceptibility, high susceptibility, and very high susceptibility. It was discovered that the current landslide areas closely match the extremely high susceptibility zone, as seen in Figure 5. This demonstrates the predictive power of the model by successfully identifying areas at high risk of landslides.

Statistical techniques can also be used to analyze the landslide susceptibility evaluation results. Together with the corresponding areas and the number of landslides linked to each susceptibility level, Table 3 lists the number of evaluation units in each of the five susceptibility levels. The landslide density within each susceptibility level was then determined by dividing the proportion of landslides by the area proportion for each level. Figure 6 presents the statistical findings.

It is evident from Table 4 and Figure 6 that the XGBoost model detects a notably higher number of landslides in the high and extremely high susceptibility zones, with a total of 905 landslides, or 76.05% of all landslides. The XGBoost model shows better predictive ability in high-risk areas, as evidenced by the 579 landslides found in the very high susceptibility zone alone, which is significantly more than the 532 landslides found by the RF model. Even though the RF model has a slightly higher landslide density than the XGBoost model, the XGBoost model displays densities of 40.72 ind/km<sup>2</sup> and 170.59 ind/km<sup>2</sup> in the high and very high susceptibility zones, respectively, while the RF model has a density of 54.40 ind/km<sup>2</sup> and 188.84 ind/km<sup>2</sup>. This discrepancy implies that the XGBoost model offers wider coverage of high-risk areas, whereas the predictions of the RF model are more concentrated within smaller area units. The RF model detected 27 landslides with a density of 1.28 ind/km<sup>2</sup>,



whereas the XGBoost model detected 17 landslides with a density of 0.79 ind/km<sup>2</sup> in the very low susceptibility zone. The XGBoost model detected 73 landslides with a density of 3.95 ind/km<sup>2</sup> in the low susceptibility zone, whereas the RF model detected 206 landslides with a density of 7.28 ind/km<sup>2</sup>. It is possible that the RF model overestimates landslides in the very low and low susceptibility zones due to the higher landslide density for these areas, which could be an indication of overprediction or misclassification. On the other hand, the XGBoost model's conservative predictions demonstrate its superior accuracy and generalization capacity. Overall, the more robust performance of the XGBoost model is demonstrated by its conservative predictions in the low susceptibility zone, which makes it more appropriate for accurate prediction in high-risk areas while avoiding over-prediction in low-risk areas. Because of this feature, the XGBoost model is better suited for a greater variety of real-world uses.

### 4.3 Model accuracy verification

As shown in Table 4 and Figure 7, model performance metrics, including accuracy, precision, recall, F1 score, KC, and MCC, were calculated along with the values of TP, FN, FP, and TN. Overall,

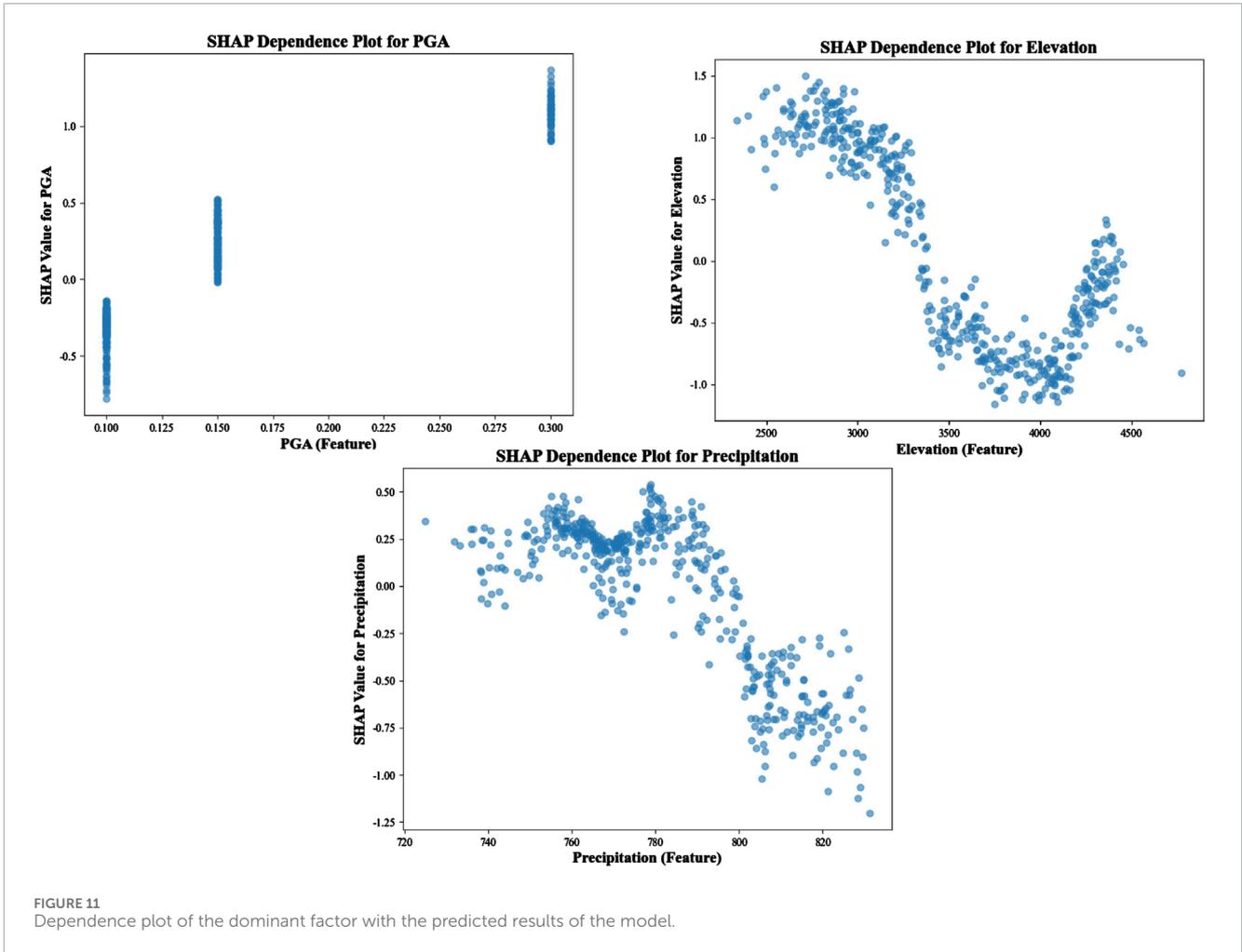
the XGBoost model outperformed the RF model in landslide susceptibility evaluation.

The ROC curves for the XGBoost and RF models are presented in Figure 8. The XGBoost model had an AUC of 0.891, slightly higher than the RF model's AUC of 0.878. Based on these results, the XGBoost model provides the best overall performance. Therefore, the evaluation results from the XGBoost model are used for further landslide susceptibility analysis and the interpretability of the evaluation factors.

## 4.4 Interpretability analysis

### 4.4.1 Importance analysis

Effective landslide disaster prevention and mitigation strategies depend on determining the main causative factors, which can be caused by several intricate factors. The 13 influencing factors that were chosen for this study have varied degrees of effect on the occurrence of landslides. These 13 factors were subjected to an importance analysis using the SHAP algorithm and the XGBoost model, as seen in Figure 9. While distance to water systems, distance to fault lines, and distance to roads have the least effect on landslide occurrence, the results show that



elevation, PGA and precipitation are the most important factors influencing landslides.

#### 4.4.2 SHAP analysis

Figure 10 displays the global SHAP explanation plot. Higher values are displayed in red, while lower values are displayed in blue. Each point in the plot represents a real sample, and its color corresponds to the magnitude of the influencing factor value. The SHAP value, which gauges each factor’s influence and contribution to the model’s anticipated result, is shown on the horizontal axis. A positive SHAP value suggests that the influencing factor significantly reduces the likelihood of landslides. The XGBoost landslide susceptibility evaluation model identifies elevation, PGA and precipitation as important triggering factors. The model’s risk prediction is positively impacted by low elevation and high PGA values, suggesting that they raise risk. The impact of precipitation on prediction is more complex, usually making a negative contribution when precipitation is high. These factors are significant in assessing the risk of geological disasters, as evidenced by their varied effects.

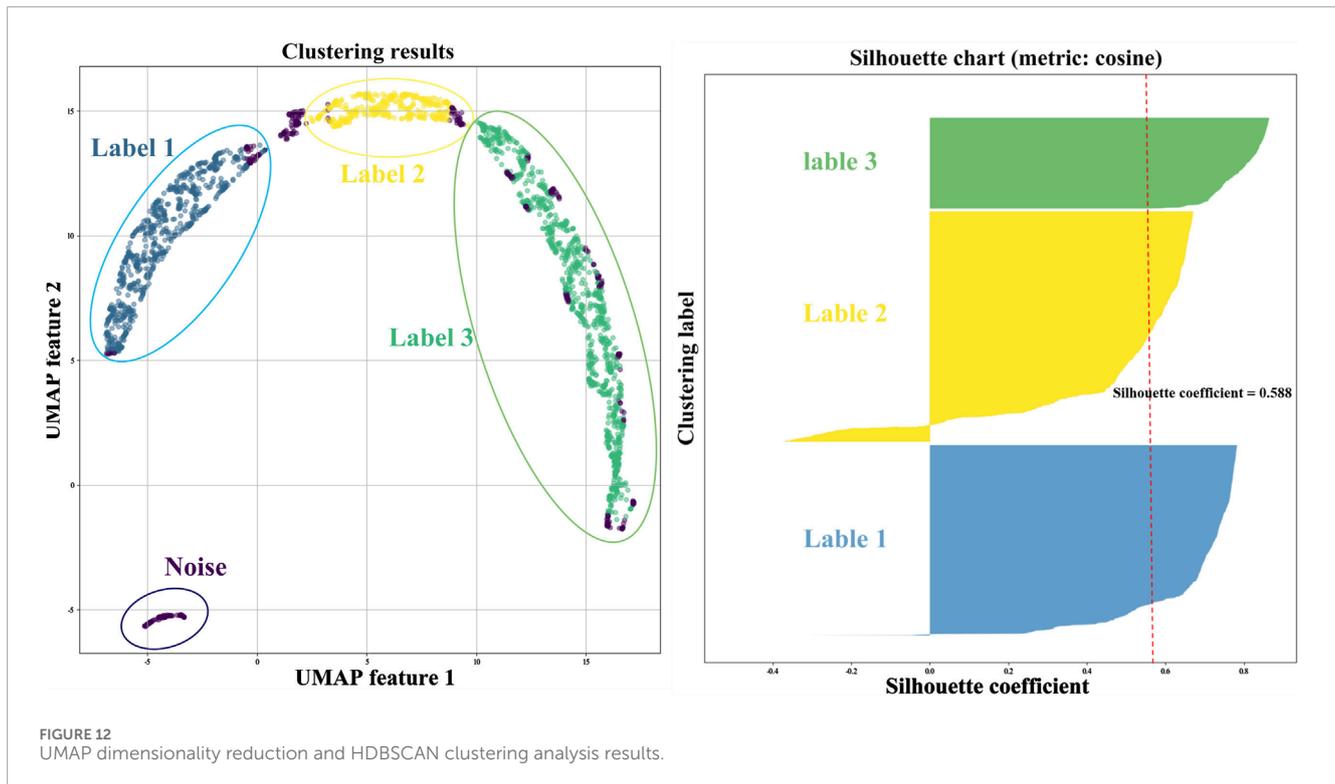
The single-factor dependence plots for the top three dominant factors based on the SHAP visualized factor importance ranking are shown in Figure 11. These plots intuitively reveal how the

different feature values of each influencing factor impact the landslide susceptibility prediction results, thereby enhancing the model’s credibility.

From the figure, elevations between 2500 m and 3400 m positively influence the model’s prediction. However, as elevation increases further, SHAP values gradually decrease, indicating a negative impact on the model. PGA values span different ranges, with SHAP values concentrated within each range. Higher PGA values lead to a significant increase in SHAP values, suggesting that higher PGA positively impacts the model’s prediction. For annual precipitation between 720 mm and 800 mm, precipitation positively affects the model’s prediction. Beyond this range, SHAP values generally decrease, indicating that higher precipitation may have a negative effect.

#### 4.4.3 UMAP and HDBSCAN analysis

In this study, we applied the UMAP and HDBSCAN models in the Python 3.10 environment. Key parameters, such as `n_neighbors` and `metric`, were tuned for the UMAP model. The `n_neighbors` parameter controls the number of nearest neighbors considered during dimensionality reduction to map the data structure, while the `metric` parameter defines the distance metric used to determine



the separation between data points. To ensure the reduced data preserved the original structure, we determined the optimal  $n_{\text{neighbors}}$  value through iterative tuning and chose the Cosine metric to measure the angle between data points in vector space. For HDBSCAN clustering, the  $\text{min\_cluster\_size}$  parameter directly influences the density distribution and granularity of clusters. After several experiments, we identified high-density landslide groups by selecting optimal clustering parameters based on data density. The results of the HDBSCAN clustering and UMAP dimensionality reduction are shown in Figure 12.

After processing with UMAP and HDBSCAN, it is evident that Barkam City's landslide points and related factors are grouped into four clusters. Some points, categorized as noise, lie in low-density areas and cannot be assigned to any identified cluster. The clustering is deemed reasonable, as indicated by the cosine metric's silhouette coefficient of 0.588.

To further analyze the interactions between different clusters in landslide susceptibility evaluation, this study used the XGBoost model to train and predict for each cluster. After training, we integrated the results of UMAP dimensionality reduction and HDBSCAN clustering to assess the importance of each evaluation factor in the XGBoost model. The results are presented in Figure 13. From the figure, it is evident that elevation, PGA and precipitation are consistently the most influential features across all clusters. Other factors, such as NDVI, distance to water systems, and lithology, show slight variations in their impact across different clusters. However, the overall ranking of the main evaluation factors remains consistent.

Specifically, elevation is the most important feature in all clusters, but its impact varies. In Label1 and Label3, the SHAP

value of elevation is negative, indicating a decrease in landslide susceptibility. In Label2, the SHAP value is positive, indicating an increase in susceptibility. PGA consistently has a positive effect on landslide susceptibility, with positive SHAP values in all clusters, meaning that higher PGA values increase susceptibility. Precipitation has a positive effect in Label1 and Label2 but a negative effect in Label3, although its impact is slightly weaker than that of elevation and PGA.

## 5 Discussion

According to an analysis of the interpretability of landslide assessment indicators, rainfall intensity, seismic activity, and topographic features are significant determinants of landslide occurrence. This outcome aligns with the theoretical comprehension of landslide formation mechanisms found in the literature currently in publication (Gaofeng et al., 2023; Hui, 2023; Jinsong et al., 2024). As shown in Figure 11, between 2,500 and 3,400 m above sea level, the slope typically increases, rainfall intensifies, and soil erosion worsens, all of which contribute to a higher landslide risk. However, landslides are less likely above 3,400 meters, where the soil layer is thin or almost nonexistent, the climate is colder, rainfall is less frequent, permafrost is more common, and bedrock is exposed. Severe seismic shaking, which greatly increases the PGA, also elevates landslide probability, particularly in the context of the Ms6.0 earthquake swarm in Barkam. Increased rainfall generally leads to soil saturation, thus raising the probability of landslides in the Barkam region, where annual rainfall ranges from 720 to 800 mm. This indicates that rainfall

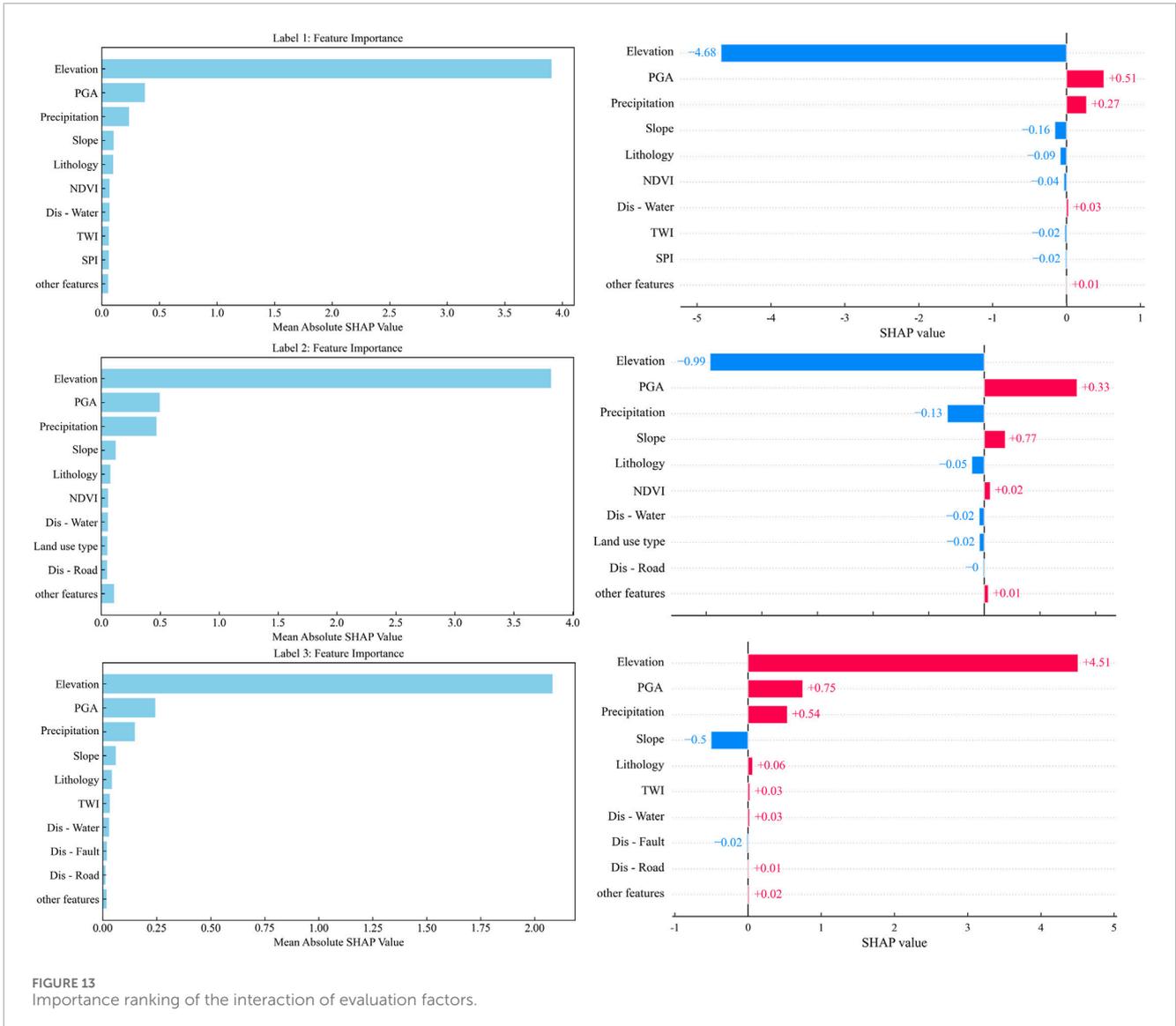


FIGURE 13 Importance ranking of the interaction of evaluation factors.

has a positive effect on landslides occurrence. Conversely, excessive rainfall can reduce landslide risk if soil permeability is low, as water infiltration becomes difficult and the amount of water that accumulates in the slope is minimal. In unsaturated soils, pore water pressure is typically low and negative, having little impact on slope stability and thereby decreasing landslide risk (Ning et al., 2018).

Machine learning models, especially RF and XGBoost, are much more flexible in handling complex datasets and do not rely on presumptions than traditional statistical methods, which makes them a significant advantage for landslide susceptibility assessment (Chen Y. et al., 2024). The “black-box” issue, which makes it challenging to directly interpret each feature’s function in the prediction process, still affects these models (Lv et al., 2024). This study combines SHAP interpretability analysis with UMAP dimensionality reduction and HDBSCAN clustering techniques to enhance model interpretability. By using the XGBoost model for one-to-many training and prediction, this approach not only

provides transparent feature difference analysis but also helps determine the significance of each assessment factor on landslide susceptibility in different clusters. The UMAP results show that the distribution patterns of various evaluation factors vary depending on the dimension. Reducing the data to two or three dimensions with UMAP unveils intricate relationships between features, offering a fresh perspective on landslide susceptibility research. In high-dimensional space, some factor interactions are not directly observable. Through HDBSCAN clustering, we identified distinct spatial heterogeneity in landslide susceptibility rather than a uniform distribution. Cluster-specific differences in susceptibility form the basis for early warning systems for landslide disasters. For example, elevation, PGA, and rainfall are dominant factors with significant impacts, as shown by the factor importance ranking in Figure 13. However, these factors contribute negatively in some clusters and positively in others. By analyzing the roles of different factors in various clusters, the study reveals the complexity of landslide occurrence.

This study has certain limitations even though the results were satisfactory. First, topography, geology, and climate are examples of significant landslide susceptibility factors that are covered by the 13 evaluation factors used in this study. However, it is possible that these factors do not adequately account for all of the intricate dynamic mechanisms influencing landslides (Song et al., 2024). For example, the landslide process could be significantly influenced by factors such as changes in groundwater levels, subtle variations in soil types, and detailed information on soil creep after earthquakes. If these factors are excluded, the model's depiction of landslide susceptibility in specific contexts may become biased. Additionally, this study applied clustering and dimensionality reduction techniques, like UMAP and HDBSCAN, which enhance the model's interpretability and work well with complex, nonlinear data. However, these techniques may still risk losing critical information as the data's dimensionality and complexity increase. Some subtle yet important features may be overlooked or oversimplified during dimensionality reduction, especially in high-dimensional data, potentially affecting how features interact with each other. Future studies could explore the integration of other dimensionality reduction methods, such as t-SNE, PCA, or factor analysis, to assess how well they simplify multi-dimensional data while preserving feature information. Moreover, incorporating automated feature engineering, advanced factor selection procedures, or combining more sophisticated data processing techniques, like deep learning methods or graph neural networks, could further improve both the model's performance and interpretability in complex scenarios.

## 6 Conclusion

- (1) The XGBoost model outperforms the RF model in predicting high and very high susceptibility areas, identifying a total of 905 landslides, which accounts for 76.05% of the total number of landslides. Among these, 579 landslides are identified in the very high susceptibility area, significantly higher than the 532 landslides identified by the RF model. Although the RF model shows higher landslide density in low and very low susceptibility areas (7.28 ind/km<sup>2</sup> in low susceptibility areas and 1.28 ind/km<sup>2</sup> in very low susceptibility areas), the XGBoost model takes a more conservative approach in these regions, identifying fewer landslides (3.95 ind/km<sup>2</sup> in low susceptibility areas and 0.79 ind/km<sup>2</sup> in very low susceptibility areas). The XGBoost model demonstrates superior performance in accurately predicting high-risk areas while avoiding over-prediction in low-risk areas, making it more suitable for a broader range of practical applications.
- (2) Based on the landslide susceptibility evaluation results, the XGBoost model outperforms the RF model in terms of performance. Specifically, the XGBoost model achieves better accuracy, precision, recall, F1 score, KC, and MCC. Additionally, its ROC curve has an AUC value of 0.891, slightly higher than the RF model's AUC of 0.878, indicating that the XGBoost model has a greater advantage in landslide susceptibility prediction.
- (3) By combining UMAP dimensionality reduction, HDBSCAN clustering, and SHAP interpretability analysis, this study

explores the key factors (elevation, PGA, precipitation, and slope) influencing landslide development and their respective effects on landslide occurrence. This approach helps to understand the underlying causes of the model's decisions, enhancing the fairness of decision-making and improving the interpretability and reliability of complex models.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Author contributions

XB: Data curation, Investigation, Software, Writing – original draft, Writing – review and editing. SF: Data curation, Methodology, Writing – original draft. ZZ: Data curation, Investigation, Writing – original draft, Writing – review and editing. KZ: Investigation, Writing – original draft. XM: Writing – review and editing.

## Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was supported by the State Grid Sichuan Electric Power Company Technology Project (52199723001D).

## Conflict of interest

Authors XB, SF, ZZ, KZ, and XM were employed by Electric Power Research Institute of State Grid Sichuan Electric Power Company.

The authors declare that this study was supported by a grant from the State Grid Electric Power Company of Sichuan Province. The involvement of the funders in this study is as follows: data acquisition, field survey support, and some experimental equipment support.

## Generative AI statement

The authors declare that no Generative AI was used in the creation of this manuscript.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

- Ahmad, S., Rizvi, Z., Arsalan Khan, M., Ahmad, J., and Wuttke, F. (2019). Experimental study of thermal performance of the backfill material around underground power cable under steady and cyclic thermal loading. *Mater. Today Proc.* 17 (P1), 85–95. doi:10.1016/j.matpr.2019.06.404
- Ahmad, S., Rizvi, Z. H., Arp, J. C. C., Wuttke, F., Tirth, V., and Islam, S. (2021). Evolution of temperature field around underground power cable for static and cyclic heating. *Energies* 14, 8191. doi:10.3390/en14238191
- Ahmad, S., Rizvi, Z. H., and Wuttke, F. (2025). Unveiling soil thermal behavior under ultra-high voltage power cable operations. *Sci. Rep.* 15, 7315. doi:10.1038/s41598-025-91831-1
- Beck-Broichsitter, S., Rizvi, Z. H., Horn, R., and Wuttke, F. (2023). Effect of gravel content on soil water retention characteristics and thermal capacity of sandy and silty soils. *Vodohospod. Cas.* 1, 71–10. doi:10.2478/johh-2023-0001
- Calderon-Guevara, W., Sanchez-Silva, M., Nitescu, B., and Villarraga, D. F. (2022). Comparative review of data-driven landslide susceptibility models: case study in the Eastern Andes mountain range of Colombia. *Nat. HAZARDS* 113, 1105–1132. doi:10.1007/s11069-022-05339-2
- Campello, R. J. G. B., Moulavi, D., and Sander, J. (2013). “Density-based clustering based on hierarchical density estimates,” in *Advances in knowledge discovery and data mining*. Editors J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu (Berlin, Heidelberg: Springer Berlin Heidelberg), 160–172.
- Chen, S., Pan, Y., Lu, C., Wang, Y., Wu, M., and Pedrycz, W. (2024). Landslide spatial prediction based on cascade forest and stacking ensemble learning algorithm. *Int. J. Syst. Sci.* 56, 658–670. doi:10.1080/00207721.2024.2408551
- Chen, Y., Li, N., Zhao, B., Xing, F., and Xiang, H. (2024). Comparison of informative modelling and machine learning methods in landslide vulnerability evaluation – a case study of Wenchuan County, China. *GEOARTO Int.* 39, 2361714. doi:10.1080/10106049.2024.2361714
- Dong, S., Liang, T., Minghui, M., and Tao, Y. (2024). Analysis of the development characteristics of co-seismic geological hazards and their controlling factors in the Maerkang Ms 6.0 earthquake swarm, Sichuan, on June 10, 2022. *J. Geomechanics* 30, 443–461. doi:10.20015/j.cnki.ISSN1000-0666.2024.0035
- Faming, H., Bin, C., Daxiong, M., and Lekai, L. (2023). Landslide susceptibility prediction modeling and interpretability based on self-screening deep learning model. *Earth Sci. Frontiers* 48, 1696–1710. doi:10.3799/dqkx.2022.247
- Gaofeng, W., Hao, L., Yuntao, T., and Zongliang, C. (2023). Study on the formation mechanism and risk prediction of high-level accumulation landslides in Bailongjiang River Basin, Gansu Province. *Chin. J. Rock Mech. Eng.* 42, 1003–1018.
- Haijun, Z., Fengshan, M., Zhiqing, L., and Jie, G. (2022). Optimization of parameters and application of probabilistic seismic LandslideHazard analysis model based on newmark displacement model: a case study in ludian earthquake area. *Earth Sci. Front.* 47, 4401–4416. doi:10.3799/dqkx.2022.289
- Han, H., Li, W., Wang, J., Qin, G., and Qin, X. (2022). Enhance explainability of manifold learning. *NEUROCOMPUTING* 500, 877–895. doi:10.1016/j.neucom.2022.05.119
- Haroon, A., Ahmad, S., and Hussain, A. (2017). “CFD prediction of loss coefficient in straight pipes,” in *Development of water resources in India*, (Cham: Springer). 477, 485. doi:10.1007/978-3-319-55125-8\_41
- Huang, F. M., Xiong, H. W., Yao, C., Catani, F., Zhou, C., and Huang, J. (2023). Uncertainties of landslide susceptibility prediction considering different landslide types. *J. ROCK Mech. Geotech.* 15, 2954–2972. doi:10.1016/j.jrmge.2023.03.001
- Huang, Y., and Zhao, L. (2018). Review on landslide susceptibility mapping using support vector machines. *CATENA* 165, 520–529. doi:10.1016/j.catena.2018.03.003
- Hui, X. (2023). Extreme characteristics and variation trend of rainfall-induced landslides in China. *Mt. Res.* 41, 545–553. doi:10.16089/j.cnki.1008-2786.000769
- Jiang, W., Li, Y., Yang, X., Deng, X., and Abbas, F. (2023). Study on landslide susceptibility in nujiang prefecture based on slope unit. *J. SOIL WATER Conserv.* 37, 160–167. doi:10.13870/j.cnki.stbcxb.2023.05.020
- Jinsong, Z., Hong, Y., Xian, Y., and Ruiming, Z. (2024). Determination of the slip surface of a thick gravel soil landslide and the prevention and control strategy. *Geol. Surv. China* 11, 108–115. doi:10.19388/j.zgdzdc.2024.02.13
- Kobak, D., and Linderman, G. C. (2021). Initialization is critical for preserving global data structure in both *t*-SNE and UMAP. *Nat. Biotechnol.* 39, 156–157. doi:10.1038/s41587-020-00809-z
- Lei, D., Ma, J., Zhang, G., Wang, Y., Deng, X., and Liu, J. (2024). Bayesian ensemble learning and Shapley additive explanations for fast estimation of slope stability with a physics-informed database. *Nat. HAZARDS* 121, 2941–2970. doi:10.1007/s11069-024-06917-2
- Li, Z., Wang, T., Zhou, Y., Liu, J., and Abbas, F. (2019). Landslide susceptibility assessment based on information value ModelLogistic regression model and their integrated model: A case in ShatangRiver basin, Qinghai Province. *Geoscience* 33, 235–245. doi:10.19657/j.geoscience.1000-8527.2019.01.23
- Lianjiang, Z., Bingxian, M., and Xuequan, Z. (2010). Clustering validity analysis based on silhouette coefficient. *Journal Computer Appl.* 30, 139–141.
- Lin, Q., Zhang, Z., Yang, Z., Zhang, X., Rong, X., Yang, S., et al. (2024). Co-seismic landslides susceptibility evaluation of Bayesian random forest considering InSAR deformation: a case study of the Luding Ms6.8 earthquake. *Geomat. Nat. HAZ RISK* 15. doi:10.1080/19475705.2024.2383783
- Lingjing, Z., Xiuzhen, L., Wenxiu, Y., and Jie, Y. (2023). Research on evaluation and prediction of potential landslide dangerous areas along the karakoram highway basin under the background of climate change. *J. Catastrophology* 38, 169–176. doi:10.3969/j.issn.1000-811X.2023.01.026
- Liu, L., and Deng, J. (2024). Multi-model combination in key steps for landslide susceptibility modeling and uncertainty analysis: a case study in Baoji City, China. *Geomat. Nat. HAZ RISK* 15. doi:10.1080/19475705.2024.2344804
- Lundberg, S. M., and Lee, S. (2017). *A unified approach to interpreting model predictions*. Red Hook, NY, USA: NIPS’17, 4768–4777.
- lv, J., Zhang, R., Shama, A., Hong, R., He, X., Wu, R., et al. (2024). Exploring the spatial patterns of landslide susceptibility assessment using interpretable Shapley method: mechanisms of landslide formation in the Sichuan-Tibet region. *J. Environ. MANAGE* 366, 121921. doi:10.1016/j.jenvman.2024.121921
- Marx, V. (2024). Seeing data as t-SNE and UMAP do. *Nat. METHODS* 21, 930–933. doi:10.1038/s41592-024-02301-x
- McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* 3, 861. doi:10.21105/joss.00861
- Ning, L., Gaunling, L., Youliang, C., and Jiancong, X. (2018). Analytic analysis on the influencing factors of Rainfall-induced shallow landslide. *J. Univ. Shanghai Sci. Technol.* 40, 65–75. doi:10.13255/j.cnki.jusst.2018.01.012
- Phuong, T. T. N., Panahi, M., Khosravi, K., Ghorbanzadeh, O., Kariminejad, N., Cerda, A., et al. (2021). Evaluation of deep learning algorithms for national scale landslide susceptibility mapping of Iran. *Geosci. Front.* 12, 505–519. doi:10.1016/j.gsf.2020.06.013
- Rizvi, Z. H., Mustafa, S. H., Ahmad, S., Sattari, A. S., Furtner, P., and Wuttke, F. (2020). “Dynamic lattice element modelling of cemented geomaterials,” in *Advances in computer methods and geomechanics*. (Singapore: Springer). doi:10.1007/978-981-15-0886-8\_53
- Sabokbar, H. F., Roodposhti, M. S., and Tazik, E. (2014). Landslide susceptibility mapping using geographically-weighted principal component analysis. *Geomorphology* 226, 15–24. doi:10.1016/j.geomorph.2014.07.026
- Song, L., Boyu, R., Mengxi, Y., and Shouhong, Z. (2024). Spatial pattern and differentiation mechanism of landslides in Beijing. A case study of “23.7” heavy rain in Mentougou district. *Sci. Soil Water Conservation* 22, 53–62. doi:10.16843/j.sswc.2024036
- Stewart, G., and Al-Khassawneh, M. (2022). An implementation of the HDBSCAN clustering algorithm. *Appl. SCI-BASEL* 12, 2405. doi:10.3390/app12052405
- Wang, D., Huang, Y., and Cai, Z. (2023). A two-phase clustering approach for traffic accident black spots identification: integrated GIS-based processing and HDBSCAN model. *Int. J. Inj. CONTROL SA30*, 270–281. doi:10.1080/17457300.2022.2164309
- Wang, H. J., Zhang, L. M., Luo, H. Y., He, J., and Cheung, R. W. M. (2021). AI-powered landslide susceptibility assessment in Hong Kong. *Eng. Geol.* 288, 106103. doi:10.1016/j.enggeo.2021.106103
- Wang, Q., Xiong, J., Cheng, W., Cui, X., and Abbas, F. (2024). Landslide susceptibility mapping methods coupling with statistical methods, machine learning models and clustering algorithms. *Journal Geo - Inf. Science* 26, 620–637. doi:10.12082/dqxxkx.2024.230427
- Weaver, C., Fortuin, A. C., Vladyka, A., and Albrecht, T. (2022). Unsupervised classification of voltammetric data beyond principal component analysis. *Chem. Commun.* 58, 10170–10173. doi:10.1039/d2cc03187f
- Xiao, X., Zou, Y., Huang, J., Luo, X., Yang, L., Li, M., et al. (2024). An interpretable model for landslide susceptibility assessment based on Optuna hyperparameter optimization and Random Forest. *Geomat. Nat. HAZ RISK* 15. doi:10.1080/19475705.2024.2347421
- Xing, F., Li, N., Zhao, B., Xiang, H., and Chen, Y. (2024). An investigation into the susceptibility to landslides using integrated learning and bayesian optimization: a case study of xichang city. *Sustainability* 16, 9085. doi:10.3390/su16209085
- Xu, Z., Che, A., and Zhou, H. (2024). Seismic landslide susceptibility assessment using principal component analysis and support vector machine. *Sci. Rep.* 14, 3734. doi:10.1038/s41598-023-48196-0
- Xuanmei, F., Xin, W., Lanxin, D., and Chengyong, F. (2022). Characteristics and spatial distribution pattern of MS6.8 luding earthquake occurred on September 5, 2022. *J. Eng. Geol.* 30, 1504–1516. doi:10.13544/j.cnki.jeg.2022-0665
- Yang, S., Tan, J., Luo, D., Wang, Y., Guo, X., Zhu, Q., et al. (2024). Sample size effects on landslide susceptibility models: a comparative study of heuristic, statistical, machine learning, deep learning and ensemble learning models with SHAP analysis. *Comput. Geosci.* 193, 105723. doi:10.1016/j.cageo.2024.105723

- Yu, B., Xing, H., Yan, J., and Li, Y. (2024). Small-scale, large impact: utilizing machine learning to assess susceptibility to urban geological disasters—a case study of urban road collapses in Hangzhou. *B Eng. Geol. Environ.* 83, 454. doi:10.1007/s10064-024-03931-3
- Yue, G., Feng, L., Ming, Z., and Peng, Y. (2024). The spatio-temporal evolution characteristics of the M 6.0 Barkam earthquake sequence in Sichuan on June 10, 2022. *Acta Seismol. Sin.*, 173–191. doi:10.11939/jass.20230104
- Zhang, J., Ma, X., Zhang, J., Sun, D., Zhou, X., Mi, C., et al. (2023). Insights into geospatial heterogeneity of landslide susceptibility based on the SHAP-XGBoost model. *J. Environ. MANAGE* 332, 117357. doi:10.1016/j.jenvman.2023.117357
- Zhang, R., Yang, Y., Wang, T., Liu, A., Lv, J., He, X., et al. (2024). Co-seismic landslide susceptibility mapping for the Luding earthquake area based on heterogeneous ensemble machine learning models. *Int. J. Digit. EARTH* 17. doi:10.1080/17538947.2024.2409337
- Zhang, X., Jiang, Y., Wang, Y., Qi, Z., and Abbas, F. (2024). Evaluation of landslide susceptibility based on multi-objective optimization method. *J. SOIL WATER Conserv.* 38, 104–112. doi:10.13870/j.cnki.stbcxb.2024.01.030
- Zhang, Z., Deng, M., Xu, S., Zhang, Y., and Abbas, F. (2022). Comparison of landslide susceptibility assessment models in zhenkang county, yunnan Province, China. *Chin. J. Rock Mech. Eng.* 41, 157–171. doi:10.13722/j.cnki.jrme.2021.0360
- Zhao, X., Chen, W., Tsangaratos, P., and Ilia, I. (2024). Evaluating landslide susceptibility: the impact of resolution and hybrid integration approaches. *Geomat. Nat. HAZ RISK* 15. doi:10.1080/19475705.2024.2409198
- Zhao, Z., Lan, H., Li, L., and Strom, A. (2024). Landslide spatial prediction using cluster analysis. *GONDWANA Res.* 130, 291–307. doi:10.1016/j.gr.2024.02.006
- Zhu, Y., Sun, D., Wen, H., Zhang, Q., Ji, Q., Li, C., et al. (2024). Considering the effect of non-landslide sample selection on landslide susceptibility assessment. *Geomat. Nat. HAZ RISK* 15. doi:10.1080/19475705.2024.2392778