

# Optimizing the Application of Machine Learning Models in Predicting Landslide Susceptibility Using the Information Value Model in Junlian County of Sichuan Basin

Lijun Qian

q1j0321@xcc.edu.cn

Xichang University

Lihua Ou

Xichang University

Guoxin Li

Xichang University

Ying Cheng

The Engineering & Technical College of Chengdu University of Technology

---

## Research Article

**Keywords:** machine learning models, landslide susceptibility prediction, information value models, non-landslide sample

**Posted Date:** July 10th, 2024

**DOI:** <https://doi.org/10.21203/rs.3.rs-4569792/v1>

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

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Optimizing the Application of Machine Learning Models in Predicting Landslide Susceptibility Using the Information Value Model in Junlian County of Sichuan Basin

Lijun Qian<sup>1</sup>, Lihua Ou<sup>1</sup>, Guoxin Li<sup>✉</sup>, Ying Cheng<sup>2</sup>,

<sup>1</sup> School of Civil and Hydraulic Engineering, Xichang University, Xichang, Sichuan 615000, China;

<sup>2</sup> The Engineering & Technical College of Chengdu University of Technology, Leshan, Sichuan 614000, China.

✉ email: ghlgx@xcc.edu.cn.

**Abstract:** Constructing accurate landslide susceptibility models is crucial for effective landslide prevention. This study explores methods to enhance the accuracy of landslide susceptibility models. This paper focuses on Junlian County, Sichuan, as the study area. Initially, a landslide inventory was created using field surveys and historical records. Eight environmental factors were identified via correlation analysis: elevation, slope, aspect, stratigraphic lithology, and distances from faults, roads, rivers, and areas of land use. Subsequently, we constructed an information value model. For training the IV-RF model, non-landslide points in areas of low susceptibility were randomly selected at various ratios (1:1, 1:2, 1:3, 1:4, 1:5). The optimal ratio was used to develop coupled models (IV-RF, IV-LR, IV-SVM, IV-BP), comparing their accuracy and discussing the impact of environmental factors on landslide susceptibility. Results indicate that: (1) the highest prediction accuracy was achieved with a non-landslide ratio of 1:1; (2) the IV-RF model achieved the highest AUC of 0.994; and (3) the most significant factors influencing landslide distribution were stratigraphic lithology and river proximity, followed by elevation and fault distance.

**Keywords:** machine learning models; landslide susceptibility prediction; information value models; non-landslide sample

## 1. Introduction

Junlian County, situated in southern Sichuan Province, lies at the intersection of the Sichuan Basin and the Yunnan-Guizhou Plateau. The region's topography and recent frequent earthquakes, combined with sudden rainfall, have led to numerous landslides. The area is experiencing rapid urban expansion, with ongoing renovation of rural homes and continuous improvements in road infrastructure. Intensified human engineering activities are progressively damaging the mountains. Consequently, accurately predicting landslides is crucial for both local residents and the government to mitigate these hazards effectively.

Landslide prediction combines complex geological engineering with both qualitative expert analysis and quantitative data-driven methods. In this context, quantitative analysis primarily employs statistical and machine learning models trained on sample data. Statistical models estimate landslide likelihood by correlating environmental factors with occurrence probabilities, using methods such as frequency ratio, information value, and weight of evidence<sup>[1, 2]</sup>. However,

statistical models struggle with non-linear variables and often fail to capture their interrelations<sup>[3]</sup>. Machine learning models, such as random forests<sup>[4, 5]</sup>, support vector machines, and BP neural networks<sup>[6]</sup>, xgboost<sup>[7]</sup>, which are widely used for predictions, can handle non-linear relationships and large datasets<sup>[8-11]</sup>. Nonetheless, these models are limited by their extensive data requirements for training, where data quality critically affects accuracy<sup>[12, 13]</sup>. Additionally, their ineffectiveness in selecting non-landslide area samples often leads to overfitting<sup>[14]</sup>. The ratio of non-landslide to landslide samples also significantly influences susceptibility assessment accuracy<sup>[15]</sup>. The opaque nature of machine learning models obscures their internal mechanisms and the significance of feature variables, complicating management in landslide-prone areas<sup>[16]</sup>. Recently, hybrid models that combine statistical and machine learning approaches have demonstrated superior accuracy in various studies<sup>[14, 17-21]</sup>.

In constructing machine learning models for landslide prediction, incorporating non-landslide samples is essential to mitigate overfitting, limit the expansion of susceptible areas, and enhance the reasonableness of predictions<sup>[15, 22, 23]</sup>. Currently, non-landslide samples lack a clear definition and universal recognition, and are typically identified as parts of non-landslide areas<sup>[15]</sup>. This ambiguity complicates the direct selection of reasonable and high-quality non-landslide samples. Selecting non-landslide samples involves considerable randomness and uncertainty<sup>[22]</sup>. Most studies only randomly select non-landslide samples once, overlooking the inherent uncertainty in this process<sup>[24]</sup>. A judicious selection of non-landslide samples can enhance model accuracy, align predictions more closely with real-world conditions, and boost model reliability<sup>[23, 25, 26]</sup>.

This study aims to develop and validate a landslide susceptibility evaluation model for Yunnan County to enhance the accuracy of landslide predictions. The research starts by conducting field surveys and organizing historical landslide data to develop a comprehensive inventory for Yunnan County. Correlation analysis is used to select low-correlation feature variables for model evaluation. Initially, an information value model constructs susceptibility zones. Then, non-landslide samples are randomly selected from these zones at ratios ranging from 1:1 to 1:5 to create the training dataset for the IV-RF model, which predicts landslide susceptibility. Optimal ratios of landslide to non-landslide samples are selected to develop IV-RF, IV-LR, IV-BP, and IV-SVM models using machine learning. These models predict landslide susceptibility and generate corresponding zoning maps. Model performance is assessed using accuracy metrics such as ACC, F1, and ROC curve AUC values to identify the most suitable landslide prediction model for Yunnan County. Finally, the significance of feature variables is assessed based on model preferences, offering a scientific foundation for future landslide risk prediction and management.

This study's significance includes: firstly, enhancing the accuracy and efficiency of landslide disaster predictions in Yunnan County, which supports disaster prevention efforts; secondly, identifying key factors influencing landslide susceptibility through model analysis, thus aiding in geological disaster risk management; and thirdly, advancing the use of machine learning in geological disaster prediction, providing a benchmark for related research.

## **2. Methods**

### **2.1 Research Technology Routes**

Figure 1 shows the analysis process flowchart:

- (1) Conduct field surveys, organize historical geological disaster records, and create a

landslide inventory for Junnan County.

(2) Perform correlation analysis on factors affecting landslide susceptibility, exclude highly correlated factors, and retain those with lower correlation for susceptibility assessment.

(3) Initially construct an information value model and delineate susceptibility areas. Randomly select non-landslide points in low susceptibility areas at ratios from 1:1 to 1:5.

(4) Use the most accurate landslide to non-landslide ratios to create a training dataset for machine learning. Develop IV-RF, IV-LR, IV-BP, and IV-SVM models for predicting landslide susceptibility and generate zoning maps.

(5) Conduct accuracy testing of the models, including ACC, F1, and AUC evaluations, to select the most suitable model for landslide prediction in this region.

(6) Calculate the importance of feature variables based on the preferences of the IV-RF, IV-LR, IV-SVM, and IV-BP models for landslide susceptibility prediction.

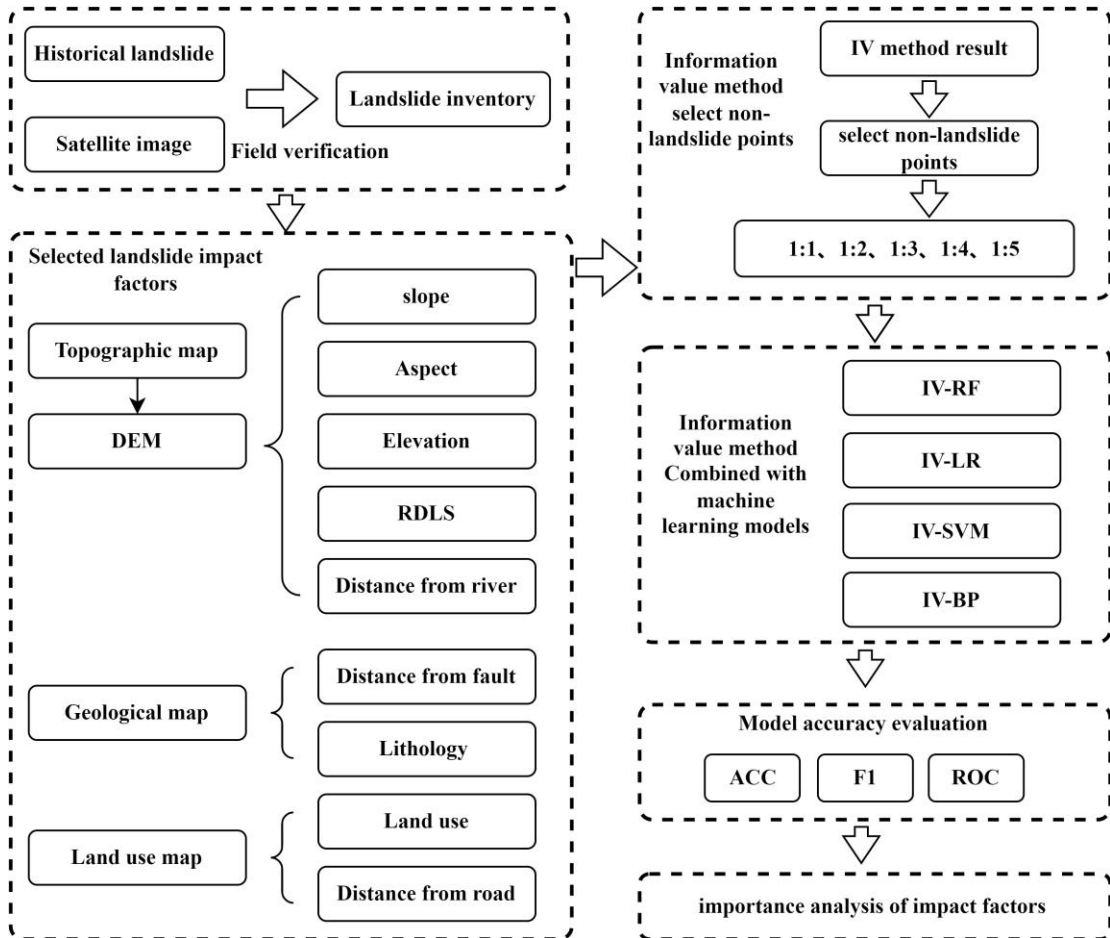


Figure 1. Flowchart of the procedures followed in this study

## 2.2 Information Value Model

Many factors influence the incidence of landslides. The Information Value Model quantifies the significance of specific factors contributing to landslide likelihood. A negligible information value indicates a minor impact on landslide likelihood. Conversely, a high information value suggests a substantial influence on landslide frequency. This model is described by Equation 1.

$$W_{Ai} = \ln \frac{N_i/S_i}{N/S} \quad (i = 1, 2, \dots, n) \quad (1)$$

$W_{Ai}$  represents the IV value for interval  $i$  of hazard element  $A$ .  $N_i$  counts the geological hazard points in interval  $i$  of factor  $A$ .  $N$  totals all hazard points across the study area.  $S_i$  measures the area of interval  $i$  of factor  $A$ , and  $S$  represents the study area's total area.

## **2.3. Machine Learning Methods**

### **2.3.1 Random Forest**

The Random Forest algorithm enhances Decision Tree methodology by orchestrating an ensemble of decision trees using feature bagging to create a robust 'forest'. The ensemble's efficacy depends on the significance of each feature, which is measured by its ability to reduce the Gini index during the tree bifurcation process<sup>[27]</sup>. Features with a greater impact on predictive accuracy are considered more important. The collective decision-making process, where different trees cast votes, gives the Random Forest model increased stability and precision against varied independent variables<sup>[28]</sup>. This collaborative approach reduces the risk of overfitting, enhancing the model's ability to generalize. In sample classification, the Random Forest model excels in efficacy and user-friendliness<sup>[29]</sup>. Its computational power makes it ideal for complex classification and regression tasks in high-dimensional datasets. The model's proficiency is demonstrated by its success in forecasting geohazard susceptibility and other applications<sup>[30, 31]</sup>.

### **2.3.2 Support Vector Machine**

Support Vector Machine (SVM) is a prominent supervised learning method designed to construct an optimal decision boundary or hyperplane. SVM increases the precision of classification or regression tasks by maximizing the margin between samples. Its exceptional performance in high-dimensional spaces makes it ideal for datasets with many feature dimensions. Additionally, SVM maintains strong performance with small-scale samples and effectively manages nonlinear relationships. In landslide susceptibility assessment, the SVM model skillfully uses limited sample data to tackle complex identification challenges.

### **2.3.3 Logistic regression**

Logistic regression is an effective algorithm for binary classification, aiming to find an optimal fitting function that accurately represents the relationship between independent and dependent variables, thus addressing binary variable issues in landslide susceptibility prediction<sup>[32]</sup>. Based on the Logit transformation, this method maps the linear combination of attributes to a probability space ranging from 0 to 1. By optimizing this mapping, logistic regression quantitatively analyzes landslide susceptibility, enabling accurate predictions<sup>[33]</sup>.

### **2.3.4 BP Neural Network**

The Back Propagation (BP) neural network is a multi-layer feedforward network trained using the error back propagation algorithm. In landslide susceptibility assessment, the BP neural network quantitatively evaluates susceptibility by learning the mapping from the input layer to the output layer. The BP neural network excels in handling nonlinear problems and mapping features. Its multi-layer structure enables it to automatically extract complex features, effectively predicting landslide susceptibility<sup>[6, 34]</sup>.

## **3. Materials**

### **3.1 study area**

Junlian County, located within Yibin City, Sichuan Province, sits at the intersection of the southern edge of the Sichuan Basin and the northern slopes of the Yungui Plateau. This strategic position serves as a natural bridge between the two distinct geological regions, with a gentle north-to-south gradient (Figure 2). The county's topography features the Daxue Mountain's main

peak in the southeast, rising to 1777.2 meters, contrasting sharply with the northern lowlands at 380 meters altitude. The county spans 48.5 kilometers east to west and 43.4 kilometers north to south, covering a total area of 1,256.42 square kilometers. Land use includes 29.4% for cultivated fields and 50.97% for forests, with the remainder allocated to construction, roads, and other infrastructure.

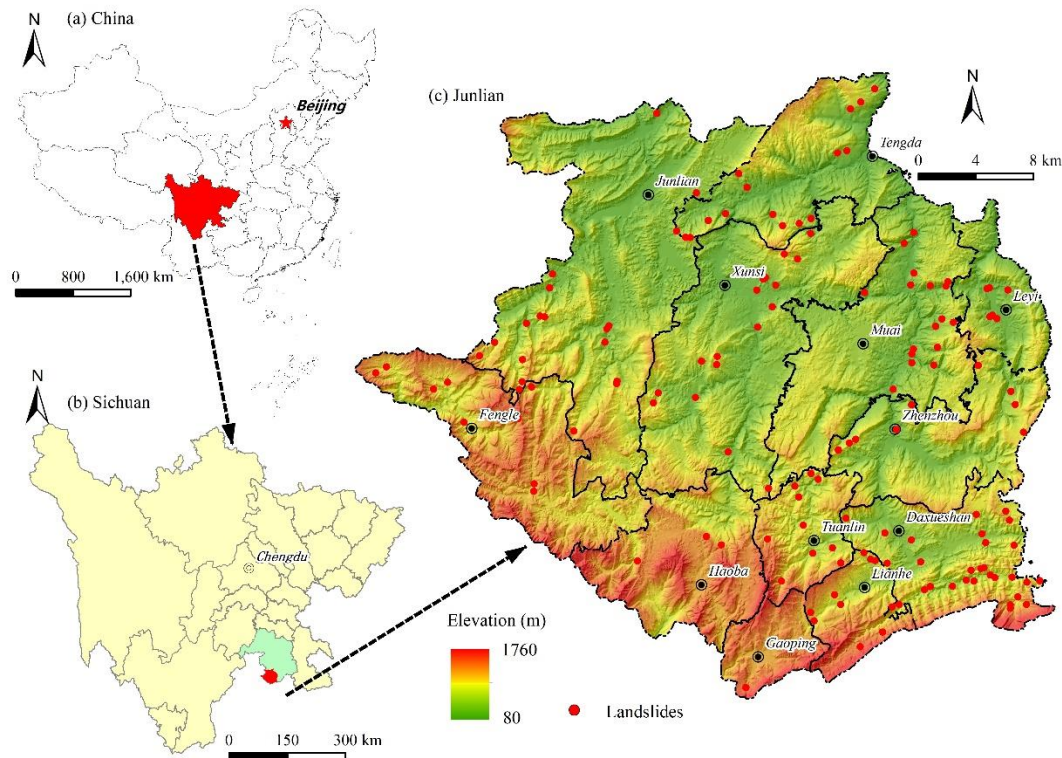


Figure 2. Maps of the study area

Junlian County enjoys a mildly warm and humid climate, with summers characterized by torrential downpours, occasional hailstorms, and strong winds. Winters feature frequent rain and fog, minimal snow, and an extended frost-free season. The county receives an average annual precipitation of 1113.6 millimeters, with 74% occurring from May to September and 54% from June to August. This heavy concentration of rainfall increases the risk of geological hazards. The county has recorded a maximum daily rainfall of 232.3 millimeters, an hourly rainfall of 123.4 millimeters, and an intense 21.15 millimeters in just 10 minutes. On average, there are five days annually with heavy rain exceeding 50 millimeters, and the longest continuous rainfall period lasts for eight days. The area boasts a network of rivers including the Zhenzhou, Jushi, Songjiang, and Leyi Rivers, which enhance the landscape and sustain the ecosystem.

Junlian County's geological profile is primarily characterized by Mesozoic Triassic, Paleozoic Permian, and Silurian formations, with additional Quaternary, Jurassic, Ordovician, and Cambrian strata. The predominant lithology features clastic rocks, supplemented by significant amounts of carbonate rocks, mudstones, and coal-bearing layers. In the southern regions, beyond the red beds, this geological composition, marked by numerous faults and fissures, results from the convergence of multiple structural elements. The county's structural grain predominantly follows east-west and northeast-southwest trends, although some areas display north-south and northwest-southeast orientations. The county lacks seismic faults and generally experiences low

seismic activity, with no recorded earthquakes exceeding a magnitude of 5.0.

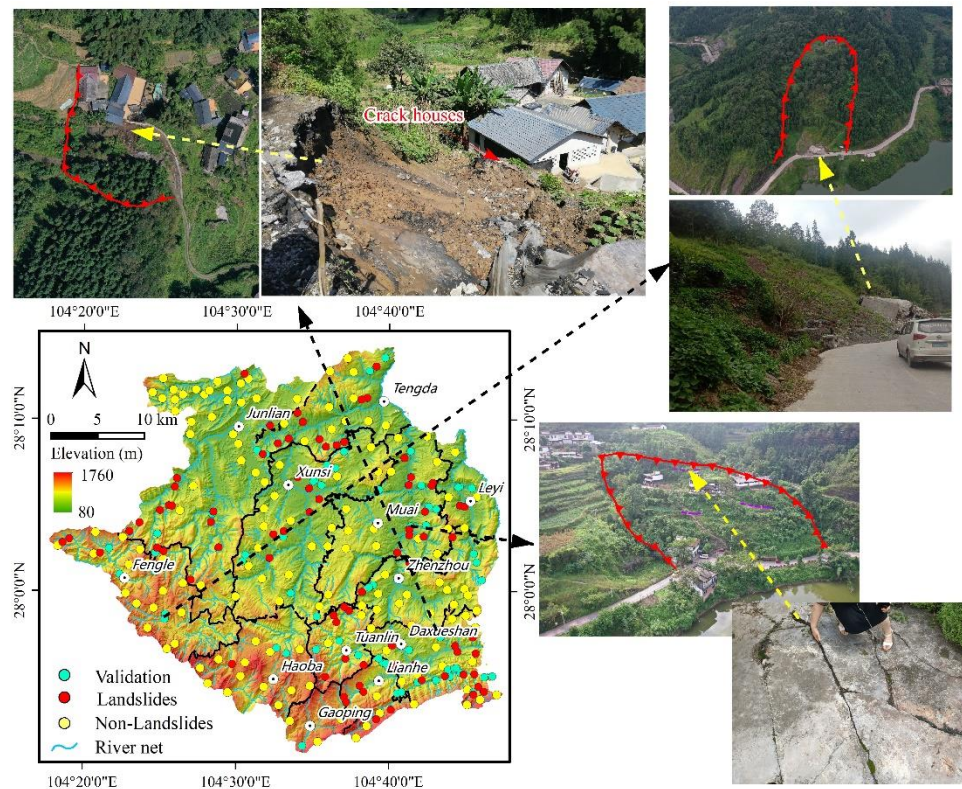


Figure 3 Distribution landslide sites and presentation of typical landslides in the study area

Landslides are common in Junlian County, with 19 events recorded between 1978 and 2020, nine of which resulted in human casualties. Most landslides in Junlian County involve soil and are relatively small (Figure 3). These landslides typically occur within the interfaces of loose Quaternary deposits or at the contact points between these deposits and bedrock. The landslide materials primarily consist of mixed sediments from multiple collapses and residual deposits on the slopes. They are loosely structured, consisting of blocky soil, gravelly soil, muddy clay, and clay. Under certain topographic conditions and rainfall, these materials are prone to cracking and sliding.

## 2.2 The Landslide Inventory

In landslide susceptibility assessments, the accuracy of the landslide inventory data is crucial for the evaluation's reliability<sup>[12, 13, 35]</sup>. This study encompasses 148 landslide locations, comprising 83 historical events and 65 sites identified by high-resolution (1-meter) Gaofen-2 satellite remote sensing, all verified by field inspections (Figure 3). The raw landslide data, presented as polygons, use the trailing edge of each landslide as the representative point for model construction. For sources and detailed explanations of data on factors influencing landslides, please see Table 1. The landslide susceptibility assessment model requires a balanced proportion of landslide and non-landslide samples. Consequently, the dataset is split into a training set (70%) and a test set (30%). Initially, the study created a landslide susceptibility map using information values, selecting an equal number of non-landslide points in non-prone areas. These points, along with field surveys, historical landslide records, and remote sensing interpretations, constitute the sample dataset.

## 2.3 Factors of landslide

Landslide occurrence results from the complex interplay of multiple factors. Currently, there is no standardized framework for defining criteria to evaluate landslide susceptibility<sup>[16]</sup>. Environmental and triggering factors are primarily considered in the occurrence of geological disasters. Extensive field surveys have led this study to identify nine key factors influencing landslides: elevation, slope gradient, aspect, topographic relief, lithology, proximity to faults, rivers, roads, and land use. Sources and detailed explanations of the data utilized in this study can be found in Table 1.

Table 1 Data sources and derived predisposing factors.

| Data sources          | scale   | Factors           | source                                                                               |
|-----------------------|---------|-------------------|--------------------------------------------------------------------------------------|
| Topographic map (DEM) | 10m     | slope             | Sichuan Provincial Bureau of Surveying,<br>Mapping and Geographic Information (2021) |
|                       |         | aspect            |                                                                                      |
|                       |         | elevation         |                                                                                      |
|                       |         | RDLS              |                                                                                      |
|                       |         | distance to river |                                                                                      |
| Geological map        | 1:50000 | distance to fault | China Geological Survey                                                              |
|                       |         | lithology         |                                                                                      |
| TM image              | 30m     | land use          | Environmental and Ecological Science Data<br>Center for West China (2022)            |

**Elevation:** The survey shows that landslides mainly occur within the 500 to 1000 meters elevation range(Figure 4a). This concentration is likely due to factors such as high population density, extensive gully development, and frequent human-induced engineering activities. Conversely, landslides are rarer in high-altitude areas, mainly due to lower human presence and activity. Elevation has been categorized into five ranges: <500m, 500-700m, 700-900m, 900-1100m, and >1100m(Figure 5a). The elevation data was sourced from the Sichuan Province Bureau of Surveying, Mapping, and Geographic Information in 2021.

**Slope:** The survey showed that landslides occur most frequently on slopes between 10 to 40 degrees, with a notable concentration of 55 landslides on slopes from 20 to 30 degrees. Above 40 degrees, the frequency of landslides significantly decreases(Figure 4b). This trend is largely due to the area's geological composition, primarily shale and mudstone, which have relatively low strength. Consequently, the slope has been categorized into six ranges: <10°, 10-20°, 20-30°, 30-40°, 40-50°, and >50°(Figure 5b).

**Aspect:** Solar radiation intensity varies with slope orientation, affecting vegetation cover, moisture evaporation, and weathering degree, which ultimately influences slope stability. Aspect has been categorized into nine orientations: flat, north, northeast, east, southeast, south, southwest, west, and northwest(Figure 4c、 5c).

**RDLS (Relative Difference of Land Surface)** measures the maximum relative height difference per unit area, indicating terrain incision within a region. Field investigations show that most landslides occur within the 5 to 15 RDLS range(Figure 4d). Landslides are rare when RDLS exceeds 20, likely due to the fragile nature of the study area's rock-soil complexes, which tend to collapse in areas with significant topographic variations. Consequently, topographic relief has been categorized into six ranges: <5°, 5-10°, 10-15°, 15-20°, 20-25°, >25°(Figure 5d). The Digital Elevation Model (DEM) data was processed with ArcGIS to calculate height differences per 100 square meters.

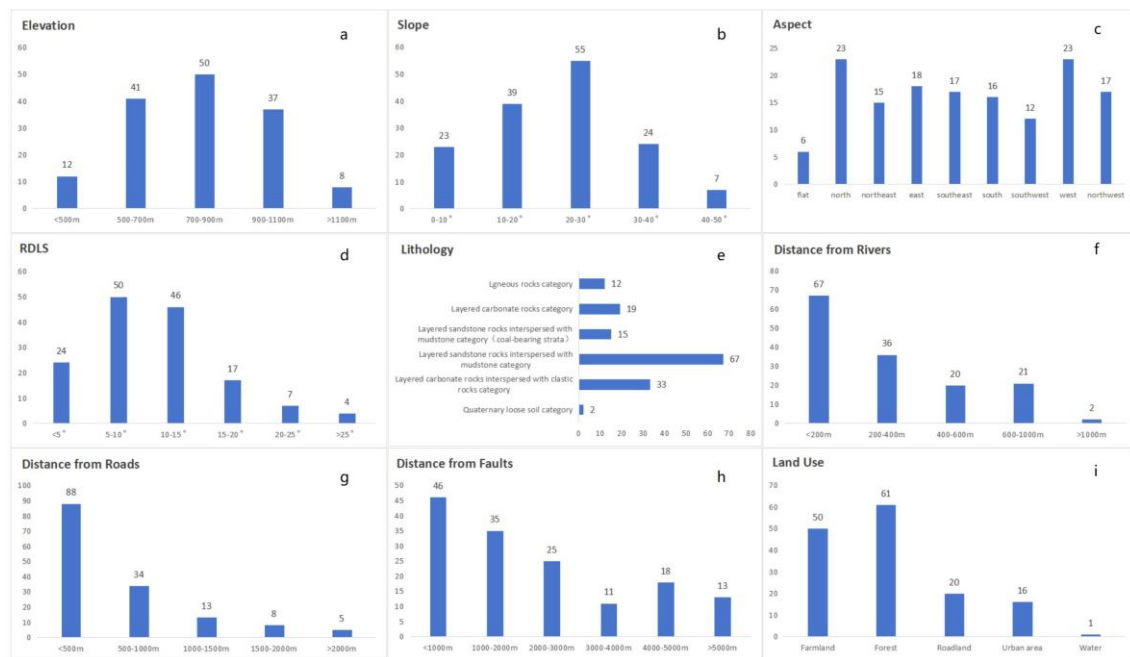


Figure 4. The relationship between the distribution of disaster points and influencing factors

**Lithology:** Stratigraphic lithology data was obtained by vectorizing the 1:50,000 geological map. Field surveys showed that landslides occurred 69 times in sandstone and mudstone interbedded with hard and semi-hard clastic rocks, and 33 times in clastic rocks and mudstone interbedded with semi-hard carbonates (Figure 4e). This indicates a significant variation in landslide probability across different rock-soil types in the area, particularly in mudstone and sandstone, which are most prone to landslides. Based on field and indoor research, lithology has been categorized into six types: Quaternary loose soils; layered carbonate rocks; carbonate rocks interspersed with clastics; igneous rocks; sandstone with mudstone; and coal-bearing sandstone with mudstone (Figure 5e).

**Distance from Rivers:** The study area features a dense river network and a well-developed hydrological system. Water infiltration softens rock layers, increasing landslide likelihood. Field surveys indicate that most landslides occur within 300 meters of rivers and water systems. According to the statistical chart, 67% of landslides occur within 300 meters of river channels (Figure 4f). As the distance from river channels increases, the incidence of landslides significantly decreases. Therefore, distances from rivers have been categorized into five ranges: <200m, 200-400m, 400-600m, 600-1000m, and >1000m (Figure 5f).

**Distance from Roads:** In mountainous terrain, road presence significantly indicates human engineering activities. Road construction often overlooks the geological environment, leading to more road-induced landslides. Geographical constraints cause most major roads to be built along mountain contours, where slope excavation at mountain bases threatens stability and can trigger landslides. Field observations reveal that areas within 100 meters of roads are most prone to landslides, recording 36 incidents, while zones within 500 meters have seen up to 88 landslides. Landslide occurrence noticeably decreases with increasing distance from roads (Figure 4g). Consequently, distances from roads are categorized into five ranges: <500m, 500-1000m, 1000-1500m, 1500-2000m, and >2000m (Figure 5g).

**Distance from Faults:** Faults significantly influence landslide occurrence. Fault development fractures rock layers, forms joints, and intensifies weathering, thus shaping landslide formation by

providing the necessary conditions and materials. Field investigations show that most landslides occur within 1000 meters of faults, with frequency noticeably decreasing as distance from faults increases(Figure 4h).. Consequently, distance from faults has been categorized into six ranges: <1000m, 1000-2000m, 2000-3000m, 3000-4000m, 4000-5000m, and >5000m(Figure 5h).

Land Use: Land use patterns directly reflect human activity intensity on slopes<sup>[36, 37]</sup>. Field surveys show that most landslides occur in forested and agricultural areas, comprising 87.3% of all landslides(Figure 4i).. Land use has been categorized into six types: agricultural land, roads, urban areas, forests, water bodies, and grasslands(Figure 5i).

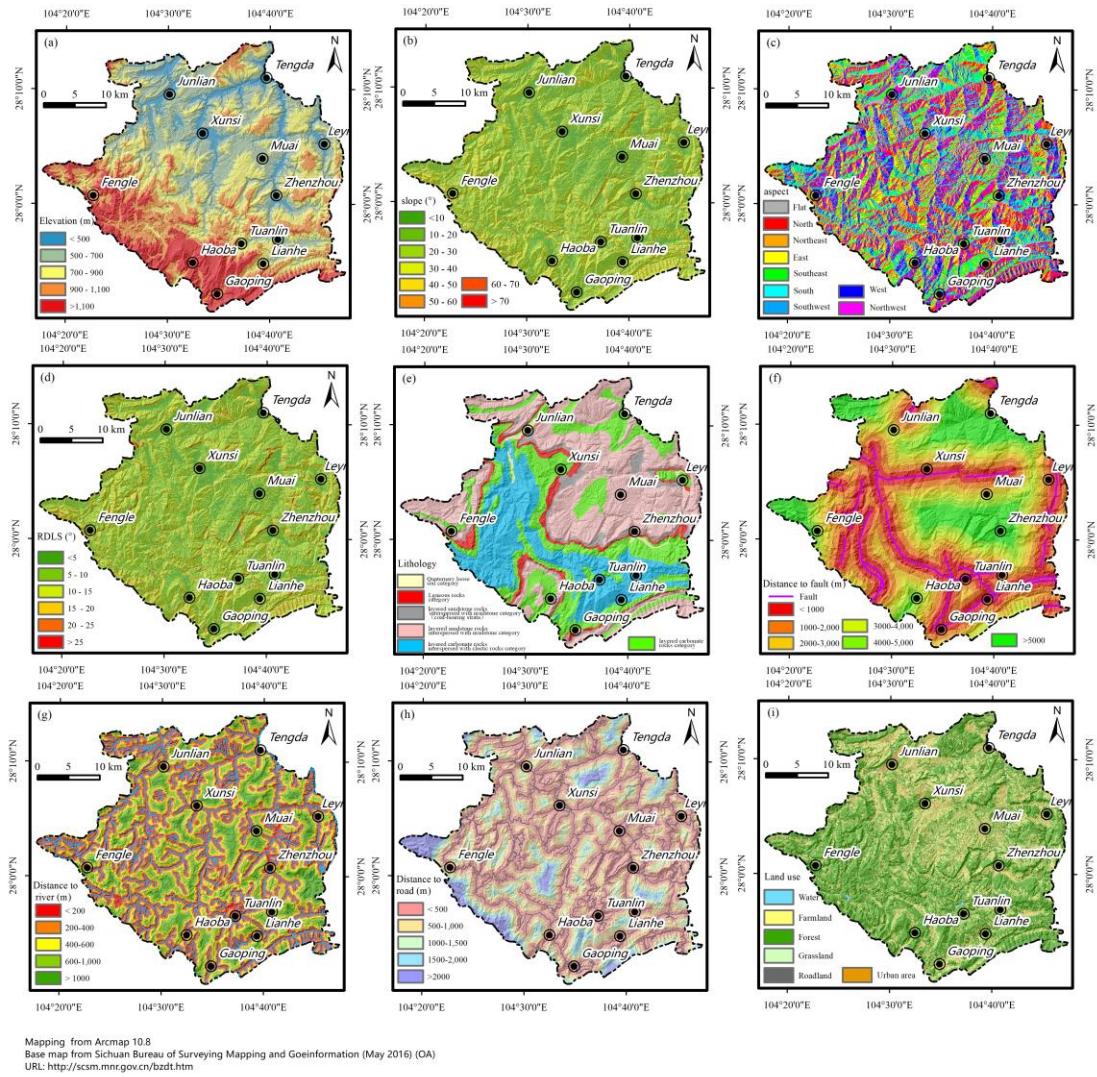


Figure 5 Grid map of the evaluation factors in the study area: (a)Elevation; (b)Slope; (c)Aspect; (d)Relief Degree of Land Surface; (e)Lithology; (f)Distance to fault; (g)Distance to river; (h)Distance to road; (i)Land Use.

### 3. Results and Discussion

#### 3.1. Multicollinearity Analysis of Conditioning Factors

Landslides are influenced by many factors, which are not entirely independent and often show correlations. In machine learning models, evaluation factors should be independent. Highly correlated factors can cause overfitting, thereby affecting prediction accuracy. This study utilized the Pearson correlation coefficient to analyze correlations among selected factors and assess their multivariate linear relationships within the model. The Pearson correlation coefficient is

defined as the covariance ratio between two variables to the product of their standard deviations. The formula for the Pearson correlation coefficient is Formula 2.

$$R = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}} \quad (2)$$

R is the Pearson correlation coefficient.  $\sigma_x$  is the standard deviation of the variable X;  $\sigma_y$  is the standard deviation of the variable Y. When  $R \geq 0.8$ , it indicates a high degree of correlation between variables. When  $0.5 \leq |R| \leq 0.8$ , it indicates a moderate level of correlation between variables. When  $0.3 \leq |R| \leq 0.5$ , it indicates a relatively low correlation between variables. When  $|R| < 0.3$ , it indicates no correlation between the two variables. According to the correlation coefficient matrix, it can be observed that the RDLS factor is significantly correlated with the slope factor (Figure 6). Therefore, the RDLS factor was excluded, and eight influencing factors—slope, elevation, aspect, engineering geological rock group, distance from the water system, distance from the road, distance from the fault, and land use—were selected as the evaluation factors for the vulnerability assessment across the entire study area.



Figure 6. The output results of Pearson correlation analysis

### 3.2 The Information Value Model and the Selection of Non-landslide Points

Selecting appropriate samples of landslide and non-landslide points is crucial for applying machine learning models to predict landslides<sup>[26]</sup>. However, no universal standard exists for selecting non-landslide samples. The study area includes 148 landslide points, which serve as positive samples, as shown in Figure 2. To improve non-landslide sample accuracy, factor correlation analysis identified eight influential factors, with their information values computed using the Information Value model. The landslide susceptibility zoning map was created using the natural breakpoint method. From low and relatively low-susceptibility zones, 148, 296, 444, 592, and 740 non-landslide points were selected, with positive to negative sample ratios of 1:1, 1:2, 1:3, 1:4, and 1:5, respectively. The distribution of these non-landslide points is illustrated in Figure 7, and the information values are detailed in Table 2.

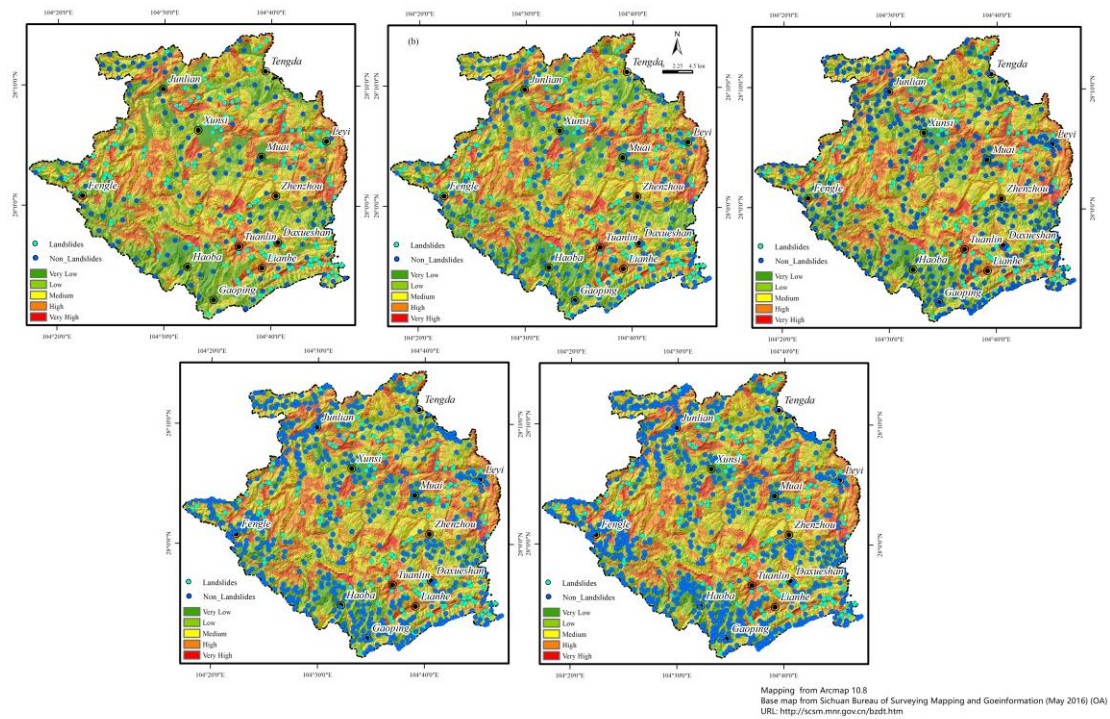


Figure 7. Information value model and selection of non-landslide points. (a) selection of non-landslide points (148);(b) selection of non-landslide points (296); (c) selection of non-landslide points (444); (d) selection of non-landslide points (592); (e) selection of non-landslide points (740).

Table 2 Information quantity for each assessment indicators

| Assessment Indicators |                                | $N_j/N$ | $S_j/S$ | $W_{ij}$ |
|-----------------------|--------------------------------|---------|---------|----------|
| Slope(° )             | <10                            | 0.1554  | 0.1960  | -0.2319  |
|                       | 10-20                          | 0.2635  | 0.2481  | 0.0603   |
|                       | 20-30                          | 0.2905  | 0.2890  | 0.0051   |
|                       | 30-40                          | 0.1419  | 0.1782  | -0.2280  |
|                       | 40-50                          | 0.0608  | 0.0703  | -0.1449  |
|                       | 50-60                          | 0.0203  | 0.0157  | 0.2565   |
|                       | 60-70                          | 0.0338  | 0.0026  | 2.5768   |
|                       | >70                            | 0.0338  | 0.0001  | 5.6292   |
| Aspect                | North                          | 0.1892  | 0.2209  | -0.1550  |
|                       | Northeast                      | 0.1081  | 0.1165  | -0.0751  |
|                       | East                           | 0.1216  | 0.1190  | 0.0222   |
|                       | Southeast                      | 0.1351  | 0.1231  | 0.0930   |
|                       | South                          | 0.1014  | 0.0931  | 0.0850   |
|                       | Southwest                      | 0.1014  | 0.0862  | 0.1621   |
|                       | West                           | 0.1149  | 0.1086  | 0.0562   |
|                       | Northwest                      | 0.1351  | 0.1326  | 0.0190   |
| Lithology             | Quaternary loose soil category | 0.0135  | 0.0088  | 0.4287   |

|                      |                                                                                         |        |        |         |
|----------------------|-----------------------------------------------------------------------------------------|--------|--------|---------|
|                      | Layered carbonate rocks category                                                        | 0.1284 | 0.2743 | -0.7592 |
|                      | Layered carbonate rocks interspersed<br>with clastic rocks category                     | 0.2230 | 0.2239 | -0.0043 |
|                      | Igneous rocks category                                                                  | 0.0811 | 0.0586 | 0.3244  |
|                      | Layered sandstone rocks interspersed<br>with mudstone category                          | 0.4527 | 0.3813 | 0.1717  |
|                      | Layered sandstone rocks interspersed<br>with mudstone category (coal-bearing<br>strata) | 0.1014 | 0.0531 | 0.6463  |
| Distance to fault(m) | <1000                                                                                   | 0.3176 | 0.2615 | 0.1943  |
|                      | 1000-2000                                                                               | 0.2365 | 0.2149 | 0.0958  |
|                      | 2000-3000                                                                               | 0.1689 | 0.1568 | 0.0748  |
|                      | 3000-4000                                                                               | 0.0811 | 0.1252 | -0.4343 |
|                      | 4000-5000                                                                               | 0.1216 | 0.0959 | 0.2379  |
|                      | >5000                                                                                   | 0.0811 | 0.1458 | -0.5869 |
| Distance to river(m) | <200                                                                                    | 0.4527 | 0.3229 | 0.3380  |
|                      | 200-400                                                                                 | 0.2432 | 0.2351 | 0.0340  |
|                      | 400-600                                                                                 | 0.1351 | 0.1690 | -0.2235 |
|                      | 600-1000                                                                                | 0.1419 | 0.1920 | -0.3023 |
|                      | >1000                                                                                   | 0.0338 | 0.0811 | -0.8753 |
| Distance to road(m)  | <500                                                                                    | 0.6014 | 0.5110 | 0.1627  |
|                      | 500-1000                                                                                | 0.2297 | 0.2681 | -0.1544 |
|                      | 1000-1500                                                                               | 0.0878 | 0.1282 | -0.3780 |
|                      | 1500-2000                                                                               | 0.0541 | 0.0520 | 0.0392  |
|                      | >2000                                                                                   | 0.0338 | 0.0407 | -0.1866 |
| Elevation(m)         | <500                                                                                    | 0.0811 | 0.0738 | 0.0945  |
|                      | 500-700                                                                                 | 0.2770 | 0.3001 | -0.0800 |
|                      | 700-900                                                                                 | 0.3378 | 0.2649 | 0.2432  |
|                      | 900-1100                                                                                | 0.2568 | 0.1990 | 0.2550  |
|                      | >1100                                                                                   | 0.0541 | 0.1622 | -1.0991 |
| Land Use             | Farmland                                                                                | 0.3851 | 0.3120 | 0.2105  |
|                      | Roadland                                                                                | 0.0270 | 0.0145 | 0.6258  |
|                      | urban area                                                                              | 0.0878 | 0.0441 | 0.6889  |
|                      | Forest                                                                                  | 0.4932 | 0.6153 | -0.2210 |
|                      | Water                                                                                   | 0.0068 | 0.0122 | -0.5946 |
|                      | Grassland                                                                               | 0.0068 | 0.0019 | 1.2747  |

### 3.3. Landslide Susceptibility Maps

Eight influential factors served as input variables for the machine learning models. Various ratios of landslide to non-landslide points were then selected as input variables. A random

sampling approach allocated 70% of the data to the training set and 30% to the testing set for constructing the IV-RF model. The natural breakpoint method, based on distinct non-landslide samples, calculated the landslide susceptibility index, delineating susceptibility levels as depicted in Figure 8. Landslide susceptibility showed high similarity across various sample sets.

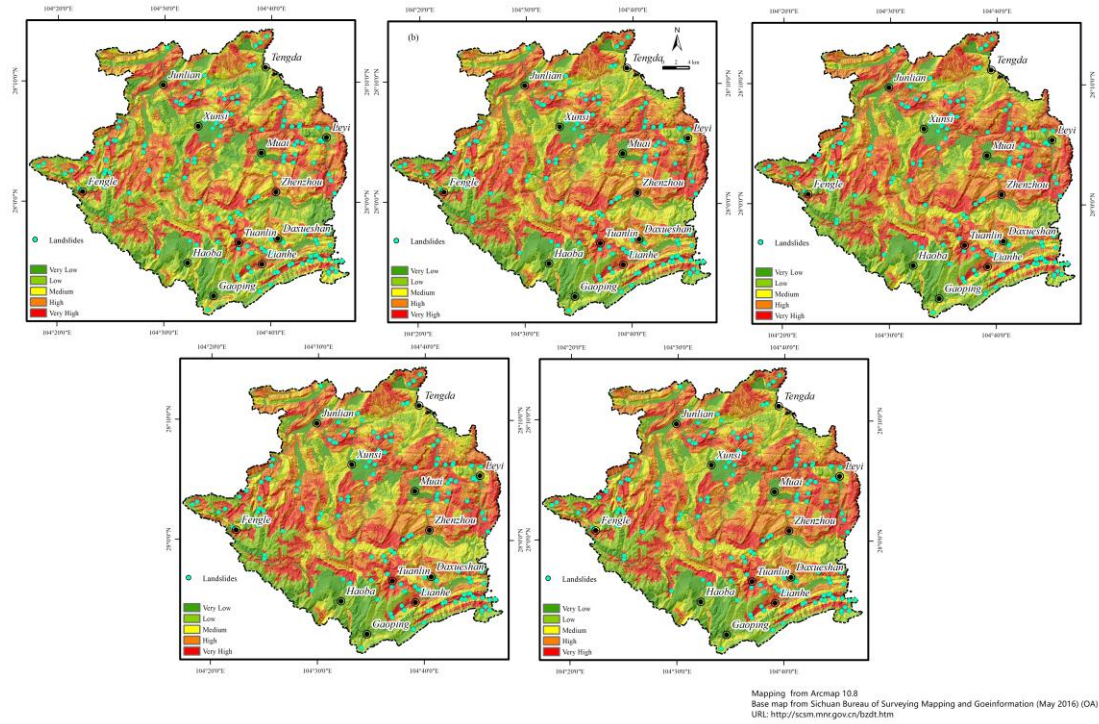


Figure 8. Landslide susceptibility evaluation results. (a) LSM of IV-RF model(148 non-landslide points);(b) LSM of IV-RF model(296 non-landslide points); (c) LSM of IV-RF model(444 non-landslide points); (d) LSM of IV-RF model(592 non-landslide points); (e) LSM of IV-RF model(740 non-landslide points).

The training set comprised 148 positive (landslide) and 148 negative (non-landslide) samples. A random sampling strategy allocated 70% of the data to the training set and 30% to the testing set. This methodology was used to construct four composite machine learning models: IV-SVM, IV-LR, IV-BP, and IV-RF, as shown in Figure 9.

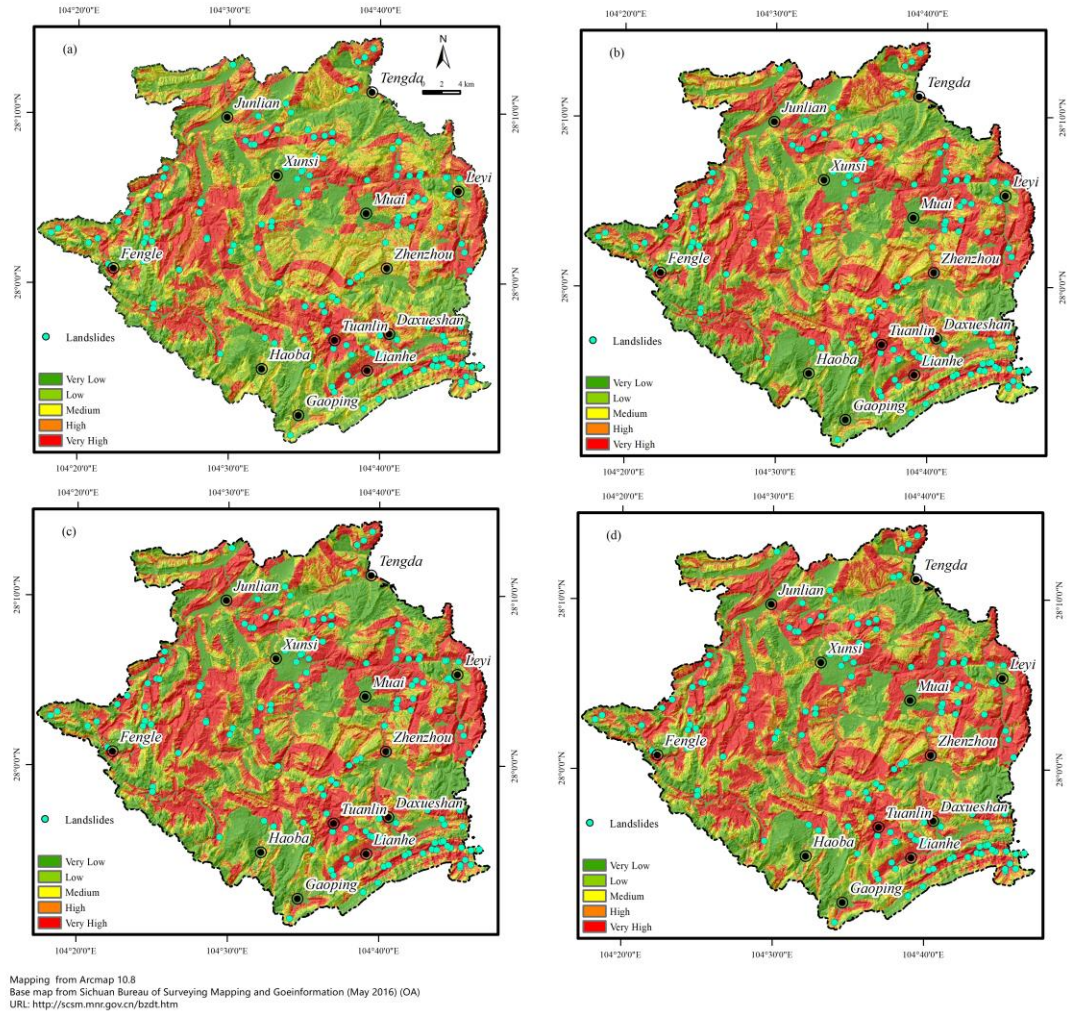


Figure 9. Landslide susceptibility evaluation results. (a) LSM of IV-RF model;(b) LSM of IV-SVM model; (c) LSM of IV-LR model; (d) LSM of IV-BP model.

Table 3. Comparison of the landslide susceptibility results using the IV-ML methods

| Method | Landslide susceptibility | Area (km <sup>2</sup> ) | Proportion of area covered (%) | number of landslides | landslides covered (%) | landslides density |
|--------|--------------------------|-------------------------|--------------------------------|----------------------|------------------------|--------------------|
| IV-RF  | very low                 | 205.06                  | 16.31%                         | 1                    | 0.68%                  | 0.00               |
|        | low                      | 224.59                  | 17.86%                         | 3                    | 2.03%                  | 0.01               |
|        | medium                   | 219.16                  | 17.43%                         | 9                    | 6.08%                  | 0.04               |
|        | high                     | 255.57                  | 20.32%                         | 37                   | 25.00%                 | 0.14               |
|        | very high                | 353.16                  | 28.08%                         | 98                   | 66.22%                 | 0.28               |
| IV-LR  | very low                 | 265.13                  | 21.08%                         | 6                    | 4.05%                  | 0.02               |
|        | low                      | 159.59                  | 12.69%                         | 3                    | 2.03%                  | 0.02               |
|        | medium                   | 142.61                  | 11.34%                         | 9                    | 6.08%                  | 0.06               |
|        | high                     | 201.73                  | 16.04%                         | 26                   | 17.57%                 | 0.13               |
|        | very high                | 488.48                  | 38.84%                         | 104                  | 70.27%                 | 0.21               |
| IV-SVM | very low                 | 239.58                  | 19.05%                         | 5                    | 3.38%                  | 0.02               |

|       |           |        |        |     |        |      |
|-------|-----------|--------|--------|-----|--------|------|
|       | low       | 173.43 | 13.79% | 1   | 0.68%  | 0.01 |
|       | medium    | 168.25 | 13.38% | 14  | 9.46%  | 0.08 |
|       | high      | 242.14 | 19.26% | 35  | 23.65% | 0.14 |
|       | very high | 434.13 | 34.52% | 93  | 62.84% | 0.21 |
| IV-BP | very low  | 234.01 | 18.61% | 5   | 3.38%  | 0.02 |
|       | low       | 161.33 | 12.83% | 2   | 1.35%  | 0.01 |
|       | medium    | 150.07 | 11.93% | 10  | 6.76%  | 0.07 |
|       | high      | 218.21 | 17.35% | 28  | 18.92% | 0.13 |
|       | very high | 493.92 | 39.28% | 103 | 69.59% | 0.21 |

Table 3 indicates that the IV-RF, IV-LR, IV-SVM, and IV-BP models predict areas highly susceptible to landslides to be 353.16 km<sup>2</sup>, 488.48 km<sup>2</sup>, 434.13 km<sup>2</sup>, and 493.92 km<sup>2</sup>, respectively. The IV-RF model has the lowest total area proportion at 28.08%, with 98 landslides occurring, representing 66.22% of all landslides and a density of 0.28 per km<sup>2</sup>. Although the IV-RF model predicted the smallest high-susceptibility area, it recorded the highest disaster density, making it the most effective of the four models.

### 3.4. Models Validation and Comparison

The accuracy of landslide susceptibility assessments directly impacts the model's reliability. Continuously optimizing hyperparameters and assessing accuracy and F1 scores can effectively improve landslide susceptibility assessments. This paper evaluates model performance using two indicators: accuracy and the area under the Receiver Operating Characteristic (ROC) curve.

In evaluating model performance for landslide susceptibility, true positives (TP) denote correctly classified landslide samples, while false positives (FP) indicate incorrectly classified ones. False negatives (FN) represent non-landslide samples classified incorrectly, and true negatives (TN) represent those classified correctly. Accuracy (ACC) is defined as the ratio of correctly classified samples to the total number of samples in the model(Formula 3).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

The F1 score, combining precision and recall, is a crucial metric for assessing classification model performance(Formula 4). An F1 score of 1 signifies perfect precision and recall, whereas a score of 0 denotes extremely poor performance in either metric. Consequently, a higher F1 score indicates a well-balanced model in terms of precision and recall.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

The Receiver Operating Characteristic (ROC) curve is a widely utilized quantitative tool for evaluating model accuracy, especially in validating landslide vulnerability assessments. On the ROC curve, the X-axis represents the True Negative Rate (TNR)—the likelihood of incorrectly predicting non-disaster points; the Y-axis represents the True Positive Rate (TPR)—the likelihood of correctly predicting disaster points(Formula 5, 6). The closer the ROC curve approaches the top left corner, the greater the accuracy of the model's predictions. The Area Under the Curve (AUC) value is commonly employed to gauge the accuracy of predictions. The AUC value denotes the area under the ROC curve, enclosed by the coordinate axes. Typically, AUC values range from 0.5

to 1, where values nearer to 1 signify more accurate predictions.

$$TNR = \frac{TN}{(TN+FN)} \quad (5)$$

$$TPR = \frac{TP}{(TP+FN)} \quad (6)$$

Table 4 shows that the ACC values for both the training and testing sets of the IV-RF, IV-LR, IV-SVM, and IV-BP models exceed 90%, with F1 scores greater than 0.9. Specifically, the IV-RF model achieved 100% accuracy and an F1 score of 1 in the training set, demonstrating its exceptional effectiveness in assessing landslide susceptibility. Figure 10a illustrates that the AUC values for the IV-RF model with non-landslide to landslide ratios of 1:2, 1:3, 1:4, and 1:5 are similar, but the 1:1 ratio has the highest AUC value. This indicates that varying the proportions of non-landslide points in non-susceptible areas does not enhance the model's accuracy. According to Figure 10b, the AUC values are 0.994 for IV-RF, 0.963 for IV-LR, 0.960 for IV-SVM, and 0.963 for IV-BP, respectively. The IV-RF model, with the highest AUC value, significantly outperforms the other models, demonstrating its superior performance and accuracy in predicting landslide susceptibility.

Table 4. Evaluation result of LSM models

| Method | Train dataset |       | Test dataset |       |
|--------|---------------|-------|--------------|-------|
|        | Accuracy      | F1    | Accuracy     | F1    |
| IV-RF  | 100%          | 1     | 93.30%       | 0.933 |
| IV-LR  | 94.20%        | 0.942 | 94.40%       | 0.944 |
| IV-SVM | 92.30%        | 0.923 | 96.60%       | 0.966 |
| IV-BP  | 92.30%        | 0.923 | 92.10%       | 0.922 |

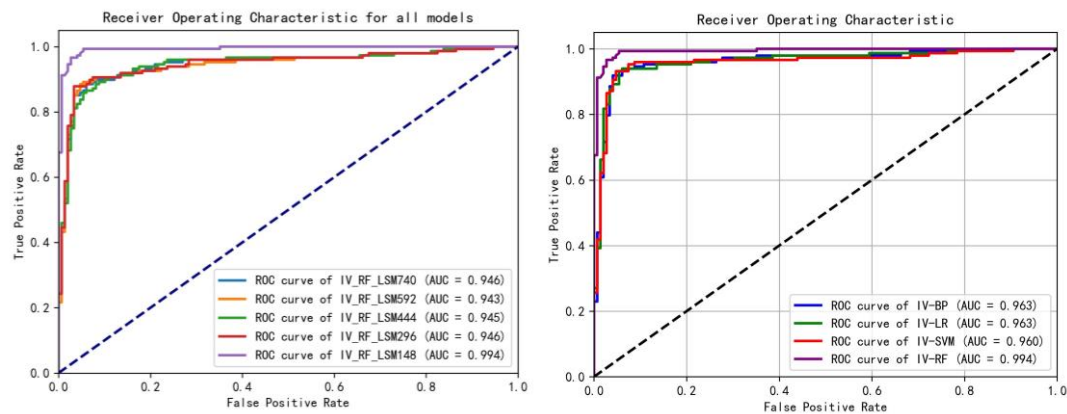


Figure 10 ROC curves for different models. (a) ROC curves for different LSM of IV-RF model (b) IV-ML models

### 3.5. Analysis of Landslide Susceptibility and Influencing Factors

Importance values of feature variables were obtained based on the propensity calculations from the IV-RF, IV-LR, IV-SVM, and IV-BP models for predicting landslide susceptibility (Figure 11). Stratigraphic lithology and proximity to rivers are identified as the most critical factors influencing landslide distribution in the study area. Elevation and proximity to faults rank as the next most significant factors. Slope, aspect, land use, and distance from roads are the least impactful factors.

Areas highly susceptible to landslides are mainly in two regions: above 700 meters elevation, within 1000 meters of faults, featuring lithologies like mudstone interbedded with carbonate rocks and sandstone (including coal-bearing strata), and within 300 meters of rivers. Specifically, these areas extend from Muhai to Leyi in the north and from Fengle to Tuanlin and Daxueshan in the south. Firstly, the region's lithology, predominantly sandstone, mudstone with clastic rocks, and clastic-carbonate rocks, is soft and brittle, increasing its susceptibility to landslides. Secondly, this area has extensively developed faults, with highly susceptible sections often forming along these faults. Additionally, the region features well-developed gullies, further contributing to its geological vulnerability to landslides. Lastly, the dense population, road construction, and housing developments involving slope cutting have continuously triggered landslides. Moderate susceptibility areas are primarily located in northern Tengda and around central Zhenzhou. These areas generally feature elevations below 900 meters, slopes under 25 degrees, brittle lithologies susceptible to weathering and fracturing, and dense road construction. Landslides primarily occur near roads and gullies. Low to very low susceptibility areas are mainly in the northern low-altitude regions, such as Yunlian Town and central Xunsi, and in the high-altitude southern regions with minimal human activity and primarily forested land use, such as the Haoba to Gaoping area.

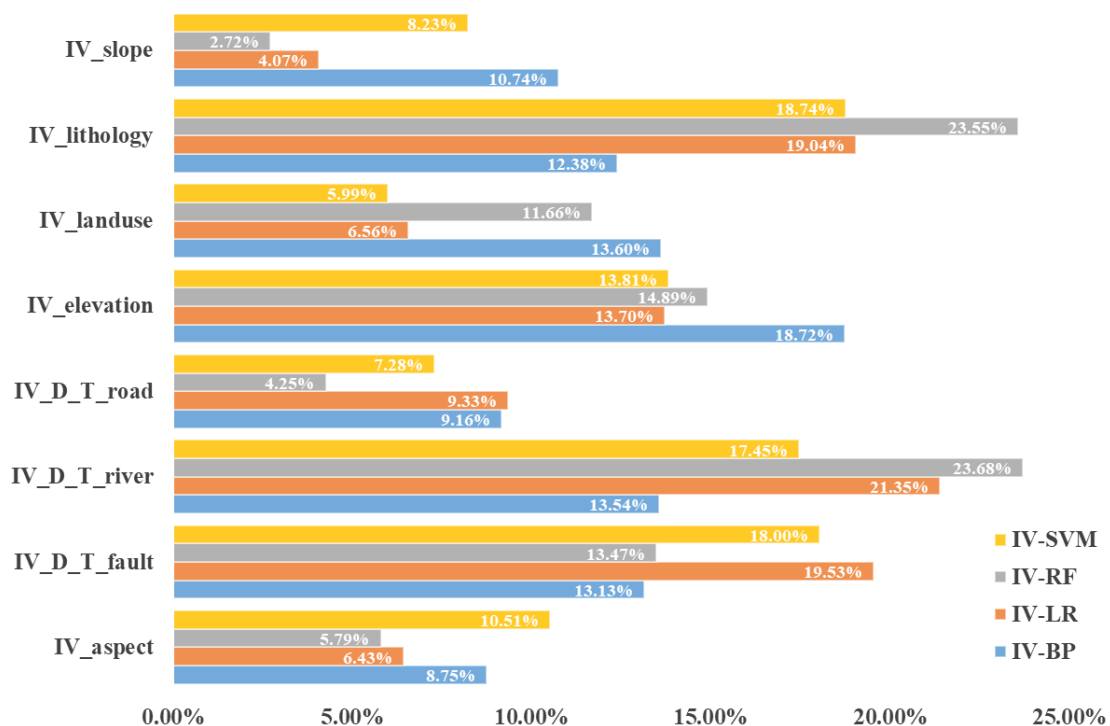


Figure 11. The importance of IV-RF model factor as calculated by the propensity score method.

#### 4. Conclusions

This study evaluates landslide susceptibility in the Junlian region, Sichuan, using various composite machine learning models. The results reveal that information value scores were calculated and the information value model was initially applied to delineate susceptibility areas. Non-landslide locations in low-susceptibility areas were randomly selected, with ratios of 1:1, 1:2, 1:3, 1:4, and 1:5 to landslide locations. The IV-RF model, once trained, was utilized to predict landslide susceptibility, revealing that varying ratios of non-landslide points minimally affected prediction outcomes. Analysis of the area under the ROC curve indicates that predictions with a

non-landslide to landslide ratio of 1:1 were the most accurate. A 1:1 non-landslide to landslide ratio was used to train and predict landslide susceptibility in the study area using four composite models: IV-RF, IV-LR, IV-SVM, and IV-BP. Accuracy tests, such as ACC, F1 scores, and area under the ROC curve, showed the IV-RF model as the most effective, achieving the highest accuracy in predicting landslide susceptibility. Propensity calculations from models IV-RF, IV-LR, IV-SVM, and IV-BP highlight stratigraphic lithology and river proximity as key factors influencing landslide distribution in the study area. Elevation and fault proximity are next in importance, whereas slope, aspect, land use, and road distance have minimal impact. This study used four machine learning models—IV-RF, IV-LR, IV-SVM, and IV-BP—to evaluate landslide susceptibility, yielding similar results. High-susceptibility areas, primarily within 300 meters of rivers, feature clastic rocks, carbonaceous mudstones, and sandstones with mudstone, including coal-bearing strata. Moderately high-susceptibility areas are mainly around the high-susceptibility zones, above 700 meters elevation and within 1000 meters of faults. Medium-susceptibility areas, surrounding moderately high zones, predominantly feature sandstone with mudstone—the most widespread lithology in the area. Low to very low susceptibility areas are mainly in the northern part, at low elevations with gentle slopes, and in the southern part, where minimal human activity, high elevations, and predominantly forested land use prevail.

### **Acknowledgements**

Elevation data were obtained from the Sichuan Provincial Bureau of Surveying, Mapping and Geographic Information (2021). This study was financially supported by the Start-up Funding for Scientific Research of the Xichang University (Grant No. YBZ202263, Grant No. YBZ202344, Grant No. YBZ202345).

### **Author contributions**

Conceptualization, L.Q., G.L. and L.O.; methodology, L.Q., Y.C.; software, L.Q., Y.C; investigation, L.Q., Y. C., G.L, L.O.and Q.L.; writing-original draft preparation, L.Q., G.L. and L.O.; writing—review and editing, L.Q., G.L.; visualization, L.Q. X.W.; supervision, L.Q., G.L.; and funding acquisition, G.L, L.O. All authors have read and agreed to the published version of the manuscript.

### **Competing interests**

The authors declare no competing interests.

### **Additional information (data availability statement)**

The datasets analyzed during the current study are not publicly available due the data are related to specific disaster coordinates; they cannot be disclosed according to Chinese law. but are available from the corresponding author on reasonable request.

### **References**

- [1] SHAHABI H, HASHIM M. Landslide susceptibility mapping using GIS-based statistical models and Remote sensing data in tropical environment [J]. Sci Rep, 2015, 5(9899).
- [2] BHAGYA S B, SUMI A S, BALAJI S, et al. Landslide Susceptibility Assessment of a Part of the Western Ghats (India) Employing the AHP and F-AHP Models and Comparison with Existing Susceptibility Maps [J]. Land, 2023, 12(2):
- [3] NWAZELIBE V E, UNIGWE C O, EGBUERI J C. Testing the performances of different fuzzy overlay methods in GIS-based landslide susceptibility mapping of Udi Province, SE Nigeria [J]. Catena, 2023, 220(

- [4] WANG Y, SUN D, WEN H, et al. Comparison of Random Forest Model and Frequency Ratio Model for Landslide Susceptibility Mapping (LSM) in Yunyang County (Chongqing, China) [J]. *Int J Environ Res Public Health*, 2020, 17(12):
- [5] ZHANG W, HE Y, WANG L, et al. Landslide Susceptibility mapping using random forest and extreme gradient boosting: A case study of Fengjie, Chongqing [J]. *Geological Journal*, 2023, 58(6): 2372-87.
- [6] HUANG J, ZENG X, DING L, et al. Landslide Susceptibility Evaluation Using Different Slope Units Based on BP Neural Network [J]. *Comput Intell Neurosci*, 2022, 2022(9923775).
- [7] CAN R, KOCAMAN S, GOKCEOGLU C. A Comprehensive Assessment of XGBoost Algorithm for Landslide Susceptibility Mapping in the Upper Basin of Ataturk Dam, Turkey [J]. *Applied Sciences*, 2021, 11(11):
- [8] HABUMUGISHA J M, CHEN N, RAHMAN M, et al. Landslide Susceptibility Mapping with Deep Learning Algorithms [J]. *Sustainability*, 2022, 14(3):
- [9] AZARAFZA M, AZARAFZA M, AKGUN H, et al. Deep learning-based landslide susceptibility mapping [J]. *Sci Rep*, 2021, 11(1): 24112.
- [10] JIANG Z, WANG M, LIU K. Comparisons of Convolutional Neural Network and Other Machine Learning Methods in Landslide Susceptibility Assessment: A Case Study in Pingwu [J]. *Remote Sensing*, 2023, 15(3):
- [11] LIU S, WANG L, ZHANG W, et al. A comprehensive review of machine learning-based methods in landslide susceptibility mapping [J]. *Geological Journal*, 2023, 58(6): 2283-301.
- [12] SAHRANE R, BOUNAB A, EL KHARIM Y. Investigating the effects of landslides inventory completeness on susceptibility mapping and frequency-area distributions: Case of Taounate province, Northern Morocco [J]. *Catena*, 2023, 220(
- [13] HAKAN T, LUIGI L. Completeness Index for Earthquake-Induced Landslide Inventories [J]. *Engineering Geology*, 2020, 264(
- [14] LIU Y, MENG Z, ZHU L, et al. Optimizing the Sample Selection of Machine Learning Models for Landslide Susceptibility Prediction Using Information Value Models in the Dabie Mountain Area of Anhui, China [J]. *Sustainability*, 2023, 15(3):
- [15] GU T, DUAN P, WANG M, et al. Effects of non-landslide sampling strategies on machine learning models in landslide susceptibility mapping [J]. *Sci Rep*, 2024, 14(1): 7201.
- [16] ZHANG J, MA X, ZHANG J, et al. Insights into geospatial heterogeneity of landslide susceptibility based on the SHAP-XGBoost model [J]. *J Environ Manage*, 2023, 332(117357).
- [17] ZHOU X, WU W, QIN Y, et al. Geoinformation-based landslide susceptibility mapping in subtropical area [J]. *Sci Rep*, 2021, 11(1): 24325.
- [18] LI Y, DENG X, JI P, et al. Evaluation of Landslide Susceptibility Based on CF-SVM in Nujiang Prefecture [J]. *Int J Environ Res Public Health*, 2022, 19(21):
- [19] GUI J, ALEJANO L R, YAO M, et al. GIS-Based Landslide Susceptibility Modeling: A Comparison between Best-First Decision Tree and Its Two Ensembles (BagBFT and RFBFT) [J]. *Remote Sensing*, 2023, 15(4):
- [20] HE W, CHEN G, ZHAO J, et al. Landslide Susceptibility Evaluation of Machine Learning Based on Information Volume and Frequency Ratio: A Case Study of Weixin County, China [J]. *Sensors (Basel)*, 2023, 23(5):
- [21] HONG H. Assessing landslide susceptibility based on hybrid Best-first decision tree with ensemble learning model [J]. *Ecological Indicators*, 2023, 147(

- [22] CHANG Z, HUANG J, HUANG F, et al. Uncertainty analysis of non-landslide sample selection in landslide susceptibility prediction using slope unit-based machine learning models [J]. *Gondwana Research*, 2023, 117(307-20).
- [23] LIU Q, TANG A, HUANG D. Exploring the uncertainty of landslide susceptibility assessment caused by the number of non-landslides [J]. *Catena*, 2023, 227(
- [24] WANG C, LIN Q, WANG L, et al. The influences of the spatial extent selection for non-landslide samples on statistical-based landslide susceptibility modelling: a case study of Anhui Province in China [J]. *Natural Hazards*, 2022, 112(3): 1967-88.
- [25] SHAO X, MA S, XU C, et al. Effects of sampling intensity and non-slide/slide sample ratio on the occurrence probability of coseismic landslides [J]. *Geomorphology*, 2020, 363(
- [26] TANG L, YU X, JIANG W, et al. Comparative study on landslide susceptibility mapping based on unbalanced sample ratio [J]. *Sci Rep*, 2023, 13(1): 5823.
- [27] ZHANG Y, WU W, QIN Y, et al. Mapping Landslide Hazard Risk Using Random Forest Algorithm in Guixi, Jiangxi, China [J]. *ISPRS International Journal of Geo-Information*, 2020, 9(11):
- [28] DANG V-H, HOANG N-D, NGUYEN L-M-D, et al. A Novel GIS-Based Random Forest Machine Algorithm for the Spatial Prediction of Shallow Landslide Susceptibility [J]. *Forests*, 2020, 11(1):
- [29] AKINCI H, KILICOGLU C, DOGAN S. Random Forest-Based Landslide Susceptibility Mapping in Coastal Regions of Artvin, Turkey [J]. *ISPRS International Journal of Geo-Information*, 2020, 9(9):
- [30] CHENG Y-S, YU T-T, SON N-T. Random Forests for Landslide Prediction in Tsengwen River Watershed, Central Taiwan [J]. *Remote Sensing*, 2021, 13(2):
- [31] WU X, SONG Y, CHEN W, et al. Analysis of Geological Hazard Susceptibility of Landslides in Muli County Based on Random Forest Algorithm [J]. *Sustainability*, 2023, 15(5):
- [32] GU T, LI J, WANG M, et al. Landslide susceptibility assessment in Zhenxiong County of China based on geographically weighted logistic regression model [J]. *Geocarto International*, 2021, 37(17): 4952-73.
- [33] CEMILOGLU A, ZHU L, MOHAMMEDNOUR A B, et al. Landslide Susceptibility Assessment for Maragheh County, Iran, Using the Logistic Regression Algorithm [J]. *Land*, 2023, 12(7):
- [34] XUE Z, FENG W, LI B, et al. Landslide susceptibility mapping based on the coupling of two correlation methods and the BP neural network model: A case study of the Baihetan Reservoir area, China [J]. *Frontiers in Environmental Science*, 2022, 10(
- [35] GUZZETTI F, MONDINI A C, CARDINALI M, et al. Landslide inventory maps: New tools for an old problem [J]. *Earth-Science Reviews*, 2012, 112(1-2): 42-66.
- [36] BOZZOLAN E, HOLCOMBE E A, PIANOSI F, et al. A mechanistic approach to include climate change and unplanned urban sprawl in landslide susceptibility maps [J]. *Sci Total Environ*, 2023, 858(Pt 1): 159412.
- [37] XIONG H, MA C, LI M, et al. Landslide susceptibility prediction considering land use change and human activity: A case study under rapid urban expansion and afforestation in China [J]. *Sci Total Environ*, 2023, 866(161430).