Journal of Civil Engineering and Urban Planning (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/jceup.2026.080106
ISSN 2616-3969 Vol. 8 Num. 1

Landslide Susceptibility Assessment in the Plateau Based on the XGBoost Model

Qiusheng Wang[1], Youwen Cui[1], Zhang Kai[2,*]

[1]Beijing University of Technology, Beijing, 100124, China
[2]China Renewable Energy Engineering Institute, Beijing, China
[*]Corresponding author

Keywords: Susceptibility Assessment; XGBoost Model; Spatial Zoning Map

Abstract: Landslides are one of the most destructive geological hazards in the Plateau, and their susceptibility assessment is of great significance for regional disaster prevention and mitigation. This study took the Great Bend region of the Yarlung Zangbo River-an area with complex terrain, frequent geological activities, and frequent landslide hazards-as the research area. Based on GIS technology and multi-source data, 12 landslide-influencing factors including elevation, slope angle, aspect, NDVI, and stratigraphic lithology were selected. After eliminating 2 factors with multicollinearity through Pearson correlation analysis and variance inflation factor test, categorical variables were processed by one-hot encoding, and a landslide susceptibility prediction model based on the eXtreme Gradient Boosting (XGBoost) algorithm was established. The model performance was evaluated by confusion matrix and ROC curve, and the factor influence mechanism was analyzed by the SHAP method, finally generating a four-level spatial zoning map of landslide susceptibility. The results show that the XGBoost model has excellent prediction performance, with an accuracy, precision, and recall rate of 0.8778 on the test set and an AUC value of 0.92, which can effectively identify high-risk areas. The extremely high susceptibility areas are mainly distributed along the main and tributary valleys of the Yarlung Zangbo River, with a landslide density of 0.430 landslides/km ²; which is highly consistent with the actual landslide distribution. The research results provide a scientific basis for the early warning and prevention of landslide hazards in the study area, and also offer reference for relevant assessments in alpine and canyon regions of Southwest China.

1. Introduction

The southwestern region of China features extremely complex geological structures, particularly in the Yarlung Zangbo River area where harsh environmental conditions, diverse landforms, and frequent geological activities have made the Daguiwan region a high-risk zone for geological disasters. Landslides, as common geological hazards, are characterized by significant damage, wide-ranging impacts, and sudden occurrence. Landslide susceptibility assessment serves as essential preparatory work before regional early warning evaluations, primarily focusing on analyzing how environmental factors influence landslide formation within specific areas. This evaluation has made crucial contributions to regional disaster prevention efforts [1].

Landslide susceptibility assessment encompasses both qualitative and quantitative evaluation methods, playing a crucial role in geological disaster risk management. Qualitative evaluation approaches primarily rely on expert experience and knowledge-based indicator selection, being knowledge and experience-driven [2]. However, these methods depend on subjective expert judgments, lack quantitative descriptions, and exhibit inconsistent evaluation methods and scoring systems among experts, limiting large-scale application [3]. With advancements in probability statistics and computer science, quantitative evaluation methods have gained increasing adoption [4], including Random Forest models [5], XGBoost models [6], logistic regression models [7], and LightGBM models [8]. Goetz et al. conducted research on landslide susceptibility assessment, comparing predictive performance between traditional statistical methods and machine learning approaches. Results demonstrated that XGBoost models achieved superior performance among all evaluated methods [9]. When dealing with training samples containing missing feature values, XGBoost can automatically learn to accurately predict data splitting directions. Leveraging its high prediction accuracy and robust stability, XGBoost has been widely applied in fields such as medical prediction and power load estimation [10].

This study establishes a landslide disaster dataset for the major bend region of the Yarlung Zangbo River and proposes a landslide susceptibility prediction model based on the Extreme Gradient Boosting (XGBoost) algorithm. Integrating ArcGIS technology, the model performs refined raster unit segmentation of the study area and systematically identifies 12 evaluation indicators from geological environment and human activity perspectives: elevation, slope gradient, slope orientation, NDVI, lithology, rainfall, land use type, distance to rivers, planar curvature, profile curvature, TWI, and terrain undulation [11]. Through XGBoost modeling of these factors, the study achieves high-precision landslide susceptibility prediction and generates susceptibility zoning maps categorized into four levels: low, medium, high, and extremely high susceptibility.

2. Project Overview and Data Processing

2.1 Project Overview

The Great Bend Region is located at the intersection of the Himalayas, Nyainqentanglha Mountains, Gangrigob Mountains, and Hengduan Mountains, featuring predominantly high mountain gorges with extremely rugged terrain. The Great Bend lies approximately at east longitude 95 ° and north latitude 29 °. This area roughly spans from Pai Town in Milin County to Motuo County in Nyingchi City.

The region is profoundly influenced by the warm and humid Indian Ocean monsoon. From May to October each year, the southwest monsoon from the Indian Ocean carries abundant moisture, which penetrates deep into the Yarlung Zangbo River Valley, resulting in abundant precipitation in the Greater Bend area. Due to significant altitude variations within this region-from low-elevation river valleys to high-altitude peaks-the climate exhibits pronounced vertical changes. The complex terrain of the Greater Bend area, featuring alternating canyon systems, mountain ranges, and river valleys, leads to diverse local microclimates.

Geological disasters primarily manifest as landslides, rockfalls, and debris flows. Landslides exhibit spatial concentration patterns, with their scale directly proportional to local topography and elevation. Greater topographic undulations correlate with larger landslide dimensions, while higher elevations also contribute to increased landslide severity. River erosion forces also induce landslides, particularly in sharp river bends where landslides predominantly occur along the valleys of the Yarlung Zangbo River and its tributaries. Proximity to rivers correlates with intensified erosion and expanded landslide scales. This study focuses on landslide analysis, with disaster point distributions illustrated in Figure 1.

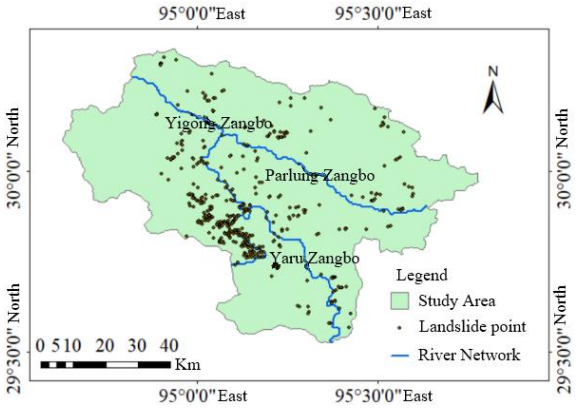

Figure 1: Spatial Distribution of Landslide Points and River Network Relationship in the Study Area

2.2 Data Processing

The landslide data were sourced from the Yajiang Large Bend Section Landslide Catalog Database (1987-2021) [12]. Elevation data were obtained from 30-meter resolution ASTER GDEM digital elevation imagery, stratigraphic lithology information from rock and soil type distribution maps at a 1:200,000 scale, rainfall data and NDVI values from the National Earth System Science Data Center [13], and land use types from Zenodo [14].

This study comprehensively considered key environmental factors influencing landslide development, selecting 12 evaluation factors including elevation, slope gradient, aspect, NDVI, lithology, rainfall, land use type, distance to rivers, planar curvature, profile curvature, TWI, and terrain undulation [15].The slope data was extracted from the Digital Elevation Model (DEM), while stratigraphic lithology and land use type data were processed through one-hot encoding for binary representation to meet model input requirements [16].A total of 898 landslide hazard sites were collected in the study area, including 449 non-slide samples (negative samples) and 449 landslide samples (positive samples). These samples were divided into a training set and a test set in an 8:2 ratio for model training and validation.

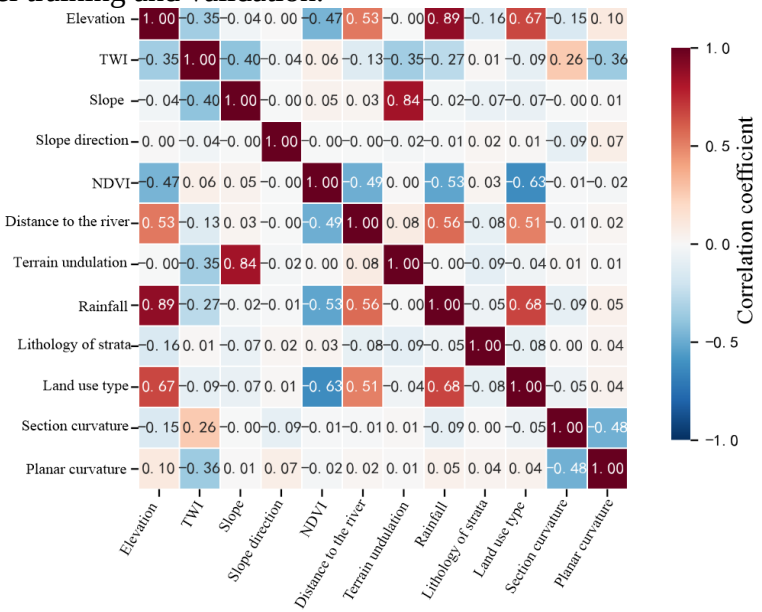

Figure 2: Pearson correlation heat map

This study imported attribute data from various classification factors into SPSS software for correlation analysis, with the resulting factor correlation coefficients presented in Figure 2. To mitigate the impact of multicollinearity among factors on model performance, Pearson correlation analysis was employed for factor screening, using a correlation coefficient threshold of |r|> 0.7 to identify significant multicollinearity. Results demonstrated an extremely strong positive correlation between slope and topographic relief (r=0.84), while correlation coefficients between rainfall and elevation (r=0.89) both exceeded the 0.7 threshold. To ensure feature independence and validity while maintaining factor representativeness, the study ultimately excluded rainfall and topographic relief factors, retaining elevation and other factors with weak or insignificant correlations for model construction.

The variance inflation factor (VIF) can be used to assess multicollinearity among factors. A VIF value greater than 5 indicates linear correlation between factors. The VIF values for each indicator factor are presented in the table 1. As shown in the table, no multicollinearity exists among the selected indicator factors in this study.

Table 1: VIF values of each indicator factor

index factor	VIF
altitude	2.482
TWI	1.763
falling gradient	1.315
aspect	1.010
NDVI	1.784
Distance to the river	1.577
stratigraphic lithology	1.043
land use type	2.435
sectional curvature	1.347
planar curvature	1.459

3. Research Methods

3.1 Principles of XGBoost Model

XGBoost (Extreme Gradient Boosting) is an ensemble learning algorithm enhanced from Gradient Boosting Decision Trees (GBDT). Its core mechanism involves sequentially constructing multiple decision trees, where each tree iteratively refines the prediction residuals of the preceding model. The final result is obtained by merging outputs from all trees, achieving both high prediction accuracy and strong generalization capabilities. Widely applied in nonlinear multi-feature problems such as landslide susceptibility assessment, the model's overall output can be expressed as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{1}$$

Here, $\hat{y}_i$ denotes the predicted value of sample i, K represents the number of decision trees, $f_k(x_i)$ indicates the prediction output of tree k for sample i, and x_i is the feature vector of sample i.

Model training focuses on minimizing the objective function, which comprises loss terms and regularization terms.

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

$\sum_{i=1}^{n} l(y_i, \hat{y}_i)$ represents the loss function, which measures the deviation between predicted values

and true values y_i. The logarithmic loss function is commonly used for classification tasks, while mean squared error is frequently employed for regression tasks; $\sum_{k=1}^{K} \Omega(f_k)$ denotes the regularization term, which controls the complexity of individual trees to prevent overfitting.

3.2 Model Training and Evaluation

This study employed grid search methodology to optimize hyperparameters of the XGBoost model, with particular emphasis on adjusting key parameters such as n estimators (number of decision trees), max depth (maximum tree depth), and learning rate to achieve optimal model performance. After model training completion, performance evaluation was conducted using confusion matrices and ROC curves. The confusion matrix visually demonstrates classification outcomes on the test set, including true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP). The ROC curve assesses overall model performance by plotting the relationship between true positive rate (TPR) and false positive rate (FPR) across different thresholds, where a larger area under the curve (AUC) indicates enhanced discriminative capability.

3.3 Landslide Susceptibility Zoning

Based on the trained XGBoost model, landslide probability prediction was conducted across the entire study area to obtain the landslide occurrence probability for each raster unit. According to the probability values, the study area was classified into four susceptibility levels: low, medium, high, and extremely high, generating a landslide susceptibility spatial zoning map that visually demonstrates the spatial distribution pattern of landslide risks [17].

4. Results and Analysis

4.1 Model Performance Evaluation

The confusion matrix of the XGBoost model on the test set shows 79 true negatives (TN), 11 false positives (FP), and 11 false negatives (FN).

The true positive (TP) rate was 79%.

Here, TP denotes a true positive instance, indicating that the actual class is positive and the model predicts it as positive; TN represents a true negative instance, indicating that the actual class is negative and the model predicts it as negative; FP refers to a false positive instance, indicating that the actual class is negative but the model predicts it as positive; FN denotes a false negative instance, indicating that the actual class is positive but the model predicts it as negative illustrated in Figure 3.

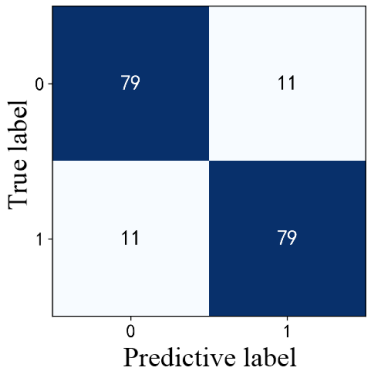

Figure 3: Confusion Matrix of the XGBoost Model

Based on the confusion matrix, the key evaluation indicators of the model are calculated as follows:

$$\text{Accuracy Rate}=\frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$\text{Precision}=\frac{TP}{TP+FP} \tag{4}$$

$$\text{Recall}=\frac{TP}{TP+FN} \tag{5}$$

The results demonstrated an accuracy rate of 87.78%, precision rate of 87.78%, and recall rate of 87.78%. The findings indicate that the XGBoost model exhibited balanced classification performance on the test set, demonstrating strong recognition capabilities for both landslide and non-slide samples without significant category bias.

The ROC curve of the XGBoost model is shown in the figure, with an area under the curve (AUC) of 0.92. An AUC value exceeding 0.9 indicates that the model possesses excellent discriminative ability, effectively distinguishing landslide samples from non-slip samples. Compared to random guessing (AUC=0.5), the XGBoost model demonstrates significant performance improvement, providing a reliable predictive foundation for landslide susceptibility assessment illustrated in Figure 4.

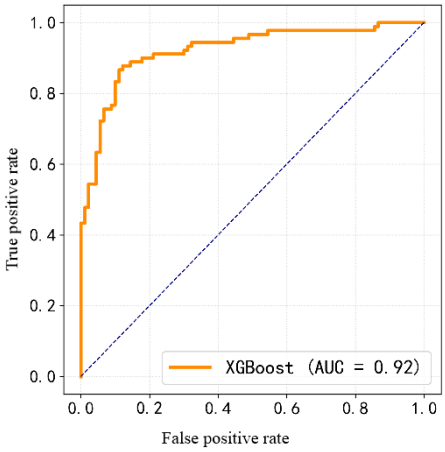

Figure 4: ROC curve of the XGBoost model

4.2 Spatial Distribution of Landslide Susceptibility

The landslide susceptibility zoning map of the study area reveals significant spatial heterogeneity in landslide susceptibility patterns illustrated in Figure 5.

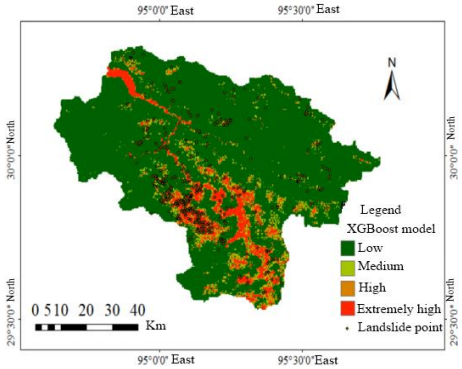

Figure 5. XGBoost Model Susceptibility Partitioning Diagram

Based on landslide occurrence probability, the area is classified into four zones: low-risk zone (0–0.25), medium-risk zone (0.25–0.5), high-risk zone (0.5–0.75), and extremely high-risk zone (0.75–1). The extremely high-risk zone primarily extends along the main course of the Yarlung Zangbo River and its tributaries, concentrated in the central and southern parts of the study area. These regions feature steep slopes and fractured strata, constituting the core landslide hazard zones. The high-risk zone predominantly surrounds the extremely high-risk area, forming linear or patchy connections with it. The medium-risk zone mainly occupies transitional areas between high-risk and low-risk zones, covering relatively smaller territories. The low-risk zone is primarily located in the northern and eastern parts of the study area, characterized by gentle terrain and lower landslide risks. Spatial distribution analysis reveals that actual landslide occurrences within the engineering zone are predominantly concentrated in extremely high-risk and high-risk zones, demonstrating strong alignment between model-predicted high-risk areas and actual landslide distributions. This confirms the model's reliability and practical applicability.

The generated landslide probability table is shown in Table 2:

Table 2: Landslide Zoning Density Table

model	probability of landslide	Area/(km²)	Number of landslide points	Landslide point density/(per km²)
XGBoost model	low	4316	186	0.043
	centre	283	48	0.170
	Gao	211	42	0.199
	polar altitude	402	173	0.430

The table 2 data reveals a clear increasing trend in landslide point density with rising susceptibility levels: low susceptibility zones show a density of 0.043 points/km²; medium susceptibility zones 0.170 points/km²; high susceptibility zones 0.199 points/km²; and extremely high susceptibility zones reaching 0.430 points/km². The density in extremely high susceptibility zones is approximately 10 times that of low susceptibility zones. Results demonstrate that the proposed XGBoost-based landslide susceptibility prediction model effectively distinguishes regions with different risk levels, with high susceptibility grades showing strong correlation with actual landslide occurrence frequencies. Although extremely high susceptibility zones account for a small geographical area, their high point density-covering 38.5% of landslide events within just 7.7% of total land area-makes them critical focal points for landslide disaster prevention. These findings provide scientific foundations for disaster mitigation strategies in the study area.

4.3 Mechanism Analysis of Landslide Susceptibility Based on SHAP

Figure 6 reveals that elevation score holds the highest importance as the core control factor for landslide development in the study area, consistent with the dominant role of topographic relief in landslide stability. River proximity ranks second, indicating significant impacts of river erosion and hydrological conditions on landslide initiation and evolution. Factors such as stratigraphic lithology, NDVI, and TWI demonstrate moderate significance, reflecting auxiliary influences from soil-rock properties, vegetation coverage, and hydrogeological conditions. Topographic factors including planar curvature, cross-sectional curvature, and slope gradient show relatively lower importance. Overall, the XGBoost model's identified key control factors align closely with the geological environment of the study area, providing scientific basis for landslide susceptibility assessment and disaster prevention mitigation.

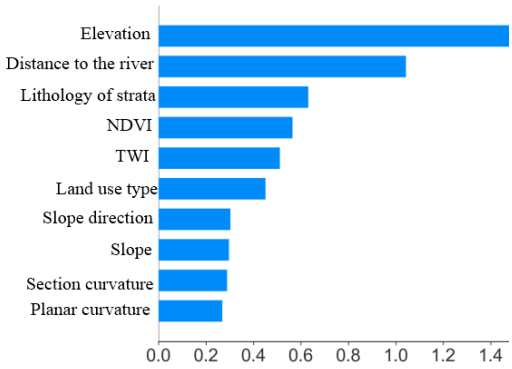

Figure 6: Feature Importance Ranking Diagram of the XGBoost Model

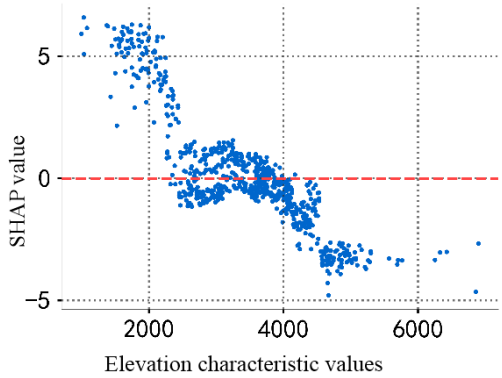

Figure 7: Elevation Feature Dependency Map

Figure 7 demonstrates that elevations below 4,000 meters facilitate landslide occurrence. This altitude range corresponds to primary human settlements and engineering development zones, where abundant precipitation increases rainfall infiltration, thereby reducing soil shear strength and elevating landslide triggering probabilities. Above 4,000 meters, landslide risks are mitigated. High-altitude regions predominantly feature gentle plateau surfaces or glacial landforms with low slope gradients. Characterized by cold climates and predominant freeze-thaw cycles, these areas exhibit relatively dense rock-soil formations. With minimal human activity and snowfall predominance, slope stability remains largely unaffected.

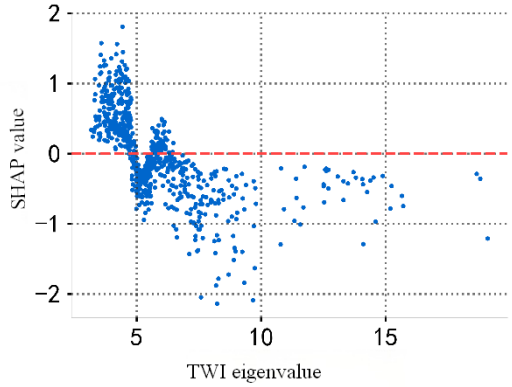

Figure 8: TWI Feature Dependency Graph

The analysis results in Figure 8 demonstrate that landslide occurrence probability increases when TWI (Twin-Water Index) is below 5, as steep slopes exhibit greater gravitational sliding components and natural slope instability conditions. Conversely, lower landslide probabilities are

observed when TWI exceeds 5. High TWI regions typically correspond to areas with gentle slopes or favorable drainage systems. Steep gradients significantly reduce sliding forces, enhancing natural slope stability. Effective drainage systems facilitate rapid water infiltration into rivers or slope discharge, thereby preventing excessive pore water pressure accumulation within slopes.

5. Conclusion

This study establishes a landslide hazard dataset for the major bend region of the Yarlung Zangbo River. A landslide susceptibility model was constructed using the XGBoost ensemble learning algorithm, with multiple collinearity factors eliminated through Pearson correlation analysis and VIF testing. Categorical variables were encoded using one-hot encoding. Model hyperparameters were optimized via grid search, and model performance was evaluated using confusion matrices and ROC curves. The study area was ultimately classified into four landslide susceptibility zones based on predicted probabilities. Key findings include:

(1) Through Pearson correlation analysis and variance inflation factor test, after excluding rainfall and topographic relief, the model achieved an accuracy, precision, and recall rate of 87.78% on the test set with an AUC value of 0.92, demonstrating balanced classification performance. This provides reliable support for regional landslide assessment.

(2) The extremely high susceptibility zone is primarily distributed along the valleys of the main and tributary streams of the Yarlung Zangbo River in the central and southern regions, with the high susceptibility zone adjacent to it. The medium and low susceptibility zones are concentrated in the gentle terrain of the northeastern region. The landslide point density in the extremely high susceptibility zone reaches 0.430 points/km², approximately 10 times that of the low susceptibility zone. The predicted results are highly consistent with the actual landslide distribution patterns.

(3) XGBoost effectively captures the nonlinear relationship between environmental factors and landslide occurrence. By combining one-hot encoding with scientific factor selection, it ensures prediction accuracy. The generated susceptibility zoning map provides scientific basis for regional landslide disaster early warning, prevention and control, as well as engineering planning. Future research can integrate dynamic triggering factors and multi-source monitoring data to construct dynamic evaluation models, further enhancing application value while offering methodological references for high mountain and canyon regions in Southwest China.

References

[1] Zhai Wenhua, Wang Xiaodong, Wu Mingtang, et al. Evaluation of geological disaster susceptibility based on frequency ratio model and random forest model coupling[J]. Journal of Natural Disaster Science, 2023,32(06):74-82.

[2] Zhang Zhipeng, Wei Zaihao. Evaluation of landslide disaster susceptibility based on weighted information quantity model: A case study of Baqiao District [J]. Science and Technology in Engineering, 2020,20(09):3492-3500.

[3] Huang F. Prediction of landslide displacement and susceptibility assessment based on 3S and artificial intelligence [D]. China University of Geosciences, 2017.

[4] Deng Shujun. Comparative study on quantitative evaluation methods for landslide susceptibility [D]. China University of Geosciences (Beijing), 2022.

[5] Liu Rui, Shi Shuxian, Sun Deliang, et al. GIS and Random Forest-Based Zoning of Landslide Susceptibility in Wushan County [J]. Journal of Chongqing Normal University (Natural Science Edition), 2020,37(03):86-96.

[6] Cui B, An Hui-Lun, Chen W-L, et al. Research on arch dam stress prediction model based on PCA-SSA-XGBoost algorithm [J]. Hydropower, 2024,50(05):45-53.

[7] Huangfu Wenchao. Application of logistic regression model in landslide disaster susceptibility assessment [D]. Donghua University of Science and Technology, 2021.

[8] Feng QR, Zhang JF, Zhu JJ, et al. Prediction method for formation fracture width based on LightGBM algorithm [J]. Energy and Environmental Protection, 2025,47(01):65-72.

[9] Goetz J N, Brenning A, Petschko H, et al. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling[J]. Computers & geosciences, 2015, 81: 1-11.

[10] Wu Hongyang, Zhou Chao, Liang Xin, et al. Evaluation and zoning of landslide susceptibility in Yanshan Township, Three Gorges Reservoir Area based on XGBoost model [J]. China Journal of Geological Hazards and Prevention, 2023,34(05):141-152.
[11] Zhao Shuai, Zhao Zhou. Evaluation of geological hazard susceptibility based on information quantity model [J]. Hydropower, 2019,45(03):27-32.
[12] Geng Haopeng. (2025). Catalog Database of Landslides in the Yajiang Large Bend Section (1987-2021). National Tibetan Plateau Science Data Center.
[13] Peng Shouzhang, Ding Yongxia, Liu Wenzhao, Li Zhi. 1 km monthly temperature and precipitation dataset for China from 1901 to 2017. Earth System Science Data, 2019, 11, 1931–1946.
[14] Jie Yang, & Xin Huang. (2026). The 30 m annual land cover datasets and its dynamics in China from 1985 to 2025 [Dataset]. Published in Earth System Science Data (1.0.5), Vol. 13, No.1, pp. 3907–3925. Zenodo.
[15] Ke C, He S, Qin Y. Comparison of natural breaks method and frequency ratio dividing attribute intervals for landslide susceptibility Mapping [J]. Bulletin of Engineering Geology and the Environment, 2023, 82(10).
[16] Agboola G, Beni L H, Elbayoumi T, et al. Optimizing landslide susceptibility mapping using machine learning and geospatial Techniques [J]. Ecological Informatics, 2024, 81: 102583.
[17] Hu Qilong, Wang Yunsheng. Evaluation of landslide susceptibility based on weighted information quantity and GIS [J]. Hydropower, 2018,44(08):31-35.