

Article

Geological Disaster Susceptibility Evaluation Using a Random Forest Empowerment Information Quantity Model

Rongwei Li ^{1,2}, Shucheng Tan ^{2,3,*} , Mingfei Zhang ⁴, Shaohan Zhang ^{1,2}, Haishan Wang ^{1,2} and Lei Zhu ⁵

¹ Institute of International Rivers and Eco-Security, Yunnan University, Kunming 650500, China; lirongweicumt@163.com (R.L.); zhangshaohan@itc.ynu.edu.cn (S.Z.); wanghaishan_vcw0@stu.ynu.edu.cn (H.W.)

² Yunnan International Joint Laboratory of Critical Mineral Resource, Kunming 650500, China

³ School of Earth Science, Yunnan University, Kunming 650500, China

⁴ Hanzhong Hydrology and Water Resources Survey Centre, Hanzhong 723000, China; zmf0624@163.com

⁵ Hubei Key Laboratory of Earthquake Early Warning, Institute of Seismology, China Earthquake Administration, Wuhan 430071, China; zhulei221@mails.ucas.ac.cn

* Correspondence: shchtan@ynu.edu.cn

Abstract: Geological hazard susceptibility assessment (GSCA) is a crucial tool widely utilized by scholars worldwide for predicting the likelihood of geological disasters. The traditional information quantity model in geological disaster susceptibility evaluation, which superimposes the information quantity of each evaluation factor without considering their weights, often negatively impacts susceptibility zoning results. This paper introduces a method employing random forest (RF) empowerment information quantity to address this issue. The method involves calculating objective weights based on a parameter-optimized random forest model, assigning these weights to each evaluation factor, and then conducting a weighted superimposition of the information. Utilizing the natural discontinuity method, the resulting comprehensive information volume map was segmented. The proposed method was applied in Kang County, Gansu Province, and its performance was compared with that of traditional methods in terms of geological disaster susceptibility zoning maps, zoning of statistical disaster point density, and receiver operating characteristic (ROC) curve accuracy. The experimental findings indicate the superior accuracy and reliability of the proposed method over the traditional approach.

Keywords: geological hazards; susceptibility; random forests; informativeness



Citation: Li, R.; Tan, S.; Zhang, M.; Zhang, S.; Wang, H.; Zhu, L. Geological Disaster Susceptibility Evaluation Using a Random Forest Empowerment Information Quantity Model. *Sustainability* **2024**, *16*, 765. <https://doi.org/10.3390/su16020765>

Academic Editor: Gwenaél Jouannic

Received: 11 December 2023

Revised: 3 January 2024

Accepted: 13 January 2024

Published: 16 January 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Geological hazards are events where the geological and natural environment deteriorates due to geological action, causing destruction of human life and property and causing serious damage to resources and the environment essential for human survival [1]. With the rapid development of human society, the utilization of natural resources to meet increasing needs has been enabled by human creativity. Nonetheless, this increased control over the environment has led to ecological damage, a product of millions of years of evolution [2]. Currently, due to accelerated human engineering activities and increased demand for natural resources, coupled with irrational land use, the frequency of geological disasters is increasing. In China, these challenges are characterized by their diversity, high frequency, wide distribution, and significant impact on human life and socio-economic activities. Annually, these disasters pose substantial threats and losses to human lives and property, impeding urban planning and construction and sustainable social development while causing extensive damage to infrastructure and the ecological environment. Slope-type geological disasters, now the second largest after earthquakes [3], include rock falls, landslides, and mudslides and occur frequently in China. A comprehensive understanding of the prevention and occurrence of geological disasters is crucial for formulating targeted

countermeasures. Thus, scientific and effective prediction of geohazard susceptibility and creation of susceptibility zoning maps are important for regional geohazard risk assessment and management. Evaluating geological disaster susceptibility offers essential guidance for risk control and prevention, aids in preventing long-term environmental damage, promotes environmental protection and ecological balance, and is crucial for rational resource use, preventing over-exploitation and deforestation, and ensuring sustainable regional development. At present, international research on geological disasters has involved the following steps: in-depth and systematic study of the disaster-causing mechanism of geological disasters from a deeper and wider perspective with the help of modern advanced scientific and technological means and methods; emphasis on the application of technical methods for disaster mapping and “3S” technology; and significant progress in the construction of regional geological disaster early warning systems and disaster management information systems in typical areas [4–6].

Currently, geological disaster susceptibility evaluation methods primarily include empirical models (fuzzy logic, hierarchical analysis, expert scoring, etc.), statistical analysis models (information quantity, frequency ratio, coefficient of determination, etc.), and machine learning models (decision tree, support vector machine, neural network, random forest, etc.) [7–9]. Among these, the empirical model relies more on experts’ experience and knowledge structures, making it susceptible to subjective factors. Consequently, the evaluation results cannot be effectively compared and analyzed [10]. The statistical analysis model necessitates independence between factors and exhibits variations in prediction results due to the absence of scientific standards or norms for continuous-type factor grading [11].

Moreover, the accuracy and applicability of empirical and statistical models are reduced when analyzing and forecasting extensive areas. With the continuous maturation and advancement of artificial intelligence algorithms, research into machine learning-based geological hazard susceptibility assessment models is becoming increasingly active. Commonly employed decision tree, support vector machine, and neural network models can better adapt to complex nonlinear features but may suffer from weak interpretability or overfitting of prediction results [12–14]. To mitigate overfitting and enhance model accuracy, integrated learning methods such as random forests have garnered widespread attention and application in geological disaster research.

For instance, Liu et al. [15] utilized the random forest algorithm to evaluate landslide susceptibility in the Wushan area of the Three Gorges Reservoir. Ahmed et al. [16] employed machine learning methods, including random forest, gradient number, multilayer perceptron, and logistic regression methods, to assess the danger of refugee camp areas in Bangladesh, with the random forest algorithm achieving the highest accuracy among the various algorithms. Merghadi et al. [17] compared the predictive capabilities of five landslide geohazard susceptibility evaluation models—random forest, gradient booster, logistic regression, neural networks, and support vector machines—in the Mira Basin of North Africa. Their findings indicated that the random forest model exhibited superior predictive performance.

In recent years, several scholars have integrated traditional models with the random forest model, leading to improved prediction accuracy and adaptability. Zheng et al. [18] applied deterministic coefficients and the random forest model in landslide susceptibility assessment in Mangshi, Yunnan Province, achieving higher accuracy than did the RF model alone. Lin et al. [19] combined the informativeness model with the random forest model in the study of landslide susceptibility in Hanzhong. They computed the weights of evaluation factors using the random forest model and then applied weighted linear combinations of the objective weights to enhance the prediction accuracy. This paper introduces an objective assignment method that combines the informativeness model and the random forest model. This method primarily determines the weights of each evaluation factor, reducing the influence of subjective factors on these weights. This approach effectively addresses the tension between the qualitative and quantitative aspects of the informativeness model and

enhances the accuracy of assessing vulnerability to geological hazards. The study area chosen is Kang County, Gansu Province, China, which is situated at the convergence of three major tectonic units. Additionally, the area is impacted by the substantial uplift of the Tibetan Plateau, which has resulted in undulating topography, active tectonic processes, and complex geological conditions.

Furthermore, human engineering activities further compound the geological challenges in this region, making it one of the most geologically active and hazard-prone areas within Gansu Province. These geological threats significantly jeopardize the safety of local residents and their property. The main purpose of this study is to provide basic information for geological disaster risk control, prevention and control, and economic and social development in Kang County and to guide relevant departments in implementing disaster prevention and mitigation strategies. Additionally, the research aims to offer valuable insights for urban spatial planning and early warning decision-making.

2. Overview of the Study Area

Kang County is situated in the southeastern part of Gansu Province, at the confluence of Shaanxi, Gansu, and Sichuan Provinces, along the upper reaches of the Jialing River and the banks of the XiHan River. Its geographic coordinates range from 105°18' to 105°50' east longitude and 32°53' to 33°39' north latitude. The county spans approximately 84.9 km from north to south and 64.2 km in width from east to west, including a total area of 2967.95 km².

The study area experiences a typical subtropical to warm temperate transition climate characterized by warm and ample rainfall. It is among the counties in Gansu with the highest precipitation levels. The average annual temperature is 11 °C, with an average annual precipitation of 777.5 mm. The precipitation distribution varies considerably across the region, generally decreasing from the southeast to the northwest. Moreover, the annual precipitation distribution is uneven, with the majority occurring between July and September, often in the form of heavy or torrential rainfall.

The study area is part of the Jialing River system and is crisscrossed by numerous perennial watercourses and streams, comprising a total of 15 major rivers. Geomorphologically, the area is characterized primarily by mountainous and river valley features. The rivers have deeply incised channels, and the uplift of bedrock in response to recent tectonic activity contributed to the undulating topography. Various geomorphological units exhibit gradient, slope, and slope direction variability, resulting in an overall terrain that is higher in the west and lower in the east.

The geological strata exposed in the region include the pre-Silurian Bikou Group, Silurian, Devonian, Carboniferous, Triassic, Jurassic, Cretaceous, Neoproterozoic, Quaternary, and magmatic rocks. Tectonic development in the area is influenced by the movement and evolution of the North China Plate, the Yangzi Plate, and the Indian Plate, making this region a significant region for plate tectonic activity in China.

The study area hosts a total of 371 disaster sites, including 186 landslides, 125 avalanches, and 60 mudslides. The geographical location and distribution of these disaster sites can be observed in Figure 1.

Geological disasters in Kang County are closely linked to rainfall and earthquakes, with most landslides occurring during the rainy season, with more than 70% occurring between July and September. In recent years, numerous geological disasters have occurred in the area, causing total economic losses of 16,112,400 RMB. For instance, the “7.17” Kang County heavy rainstorm in 2009 triggered hundreds of landslides and avalanches, cumulatively amounting to 3 million m³, blocking highways for two months and destroying extensive forest areas. Given its rich mineral resources, deep history, comprehensive infrastructure, 74% vegetation coverage, and status as an ecotourism city, conducting a geological disaster susceptibility evaluation in Kang County is critical. Although several geologists have carried out some work on mineral geology, hydrogeology, engineering geology, and disaster geology in Kang County since the 1950s, this work is still insufficient to meet certain needs; therefore, this study further identifies the types, scales, distribution

patterns and developmental characteristics of geological disasters in the study area and evaluates susceptibility to geological disasters on the basis of the collection of detailed investigations and zoning of geological disasters that have already been carried out in the survey area as well as information on the results that have already been generated by others.

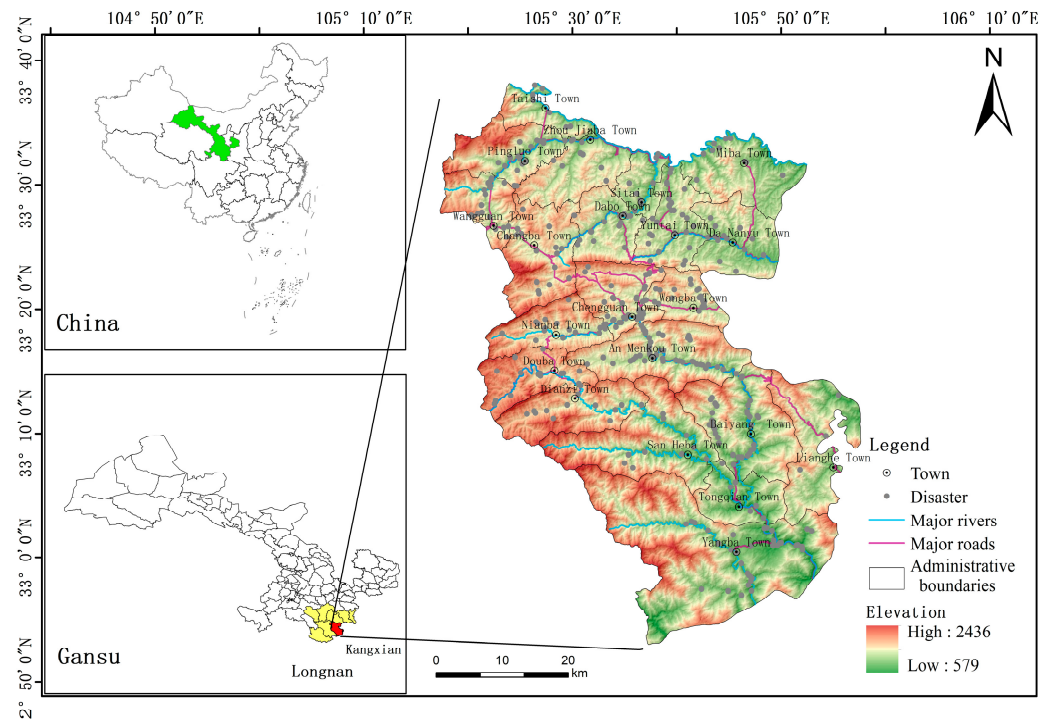


Figure 1. Overview of the study area.

3. Research Methods and Data Sources

3.1. Research Methodology

3.1.1. Information Quantity Model (IM)

The IM quantifies the informativeness values of various subcategories of each evaluation factor using statistical methods, indicating their contribution to disaster development. This objective, efficient method is widely applied in geological disaster susceptibility evaluation. The model utilizes the size and integrated level of the informativeness values contributed by each factor as standards for regional disaster susceptibility zoning and calculates the contribution of each factor to disaster informativeness as the model's core [20]. Its formula is:

(1) The evaluation indicators used to determine geological hazards in the study area are as follows:

$$I(x_i, D) = \ln \frac{S_i / A_i}{S / A} \quad (1)$$

where (x_i, D) represents the information provided by evaluation indicator x_i on geological hazard risk in the study area; S_i is the number of geohazard sites in evaluation indicator x_i ; A_i denotes the area of indicator x_i ; S is the total number of hazard sites; and A is the total area of the study area.

(2) The comprehensive information on geological hazards in the study area is as follows:

$$I = \sum_i^n I(x_i, D) = \sum_i^n \ln \frac{S_i / A_i}{S / A} \quad (2)$$

where n is the number of evaluation indicators and I represents the comprehensive information amount, with a larger value of I indicating higher geological hazard risk.

3.1.2. Random Forest (RF) Model

The RF algorithm, an integration method for decision trees, was first proposed by Breiman [21] and is a prominent algorithm of Bagging integration methods. Independent and identically distributed random vectors are used as parameter sets, and the optimal classification result is determined by voting from each decision tree model based on the independent variable X . In the RF model, training samples for each tree and split attributes for nodes are randomly selected, reducing model overfitting and enhancing robustness. Numerous theoretical and applied studies have verified the accuracy of the model, which tolerates outliers and noise in the dataset well, making it one of the best machine learning models [22–24]. These are the main four steps of the random forest algorithm:

- (1) Repeated sampling was performed from the original samples in a sample-and-return manner, leading to the formation of a training set on each occasion.
- (2) Subsequent to each sampling for training set creation, a decision tree was generated accordingly. It was assumed that, within the sample, M features were present, and n features were randomly chosen from among the M features to serve as the set of split features for every internal node of the decision tree.
- (3) The optimal method for splitting the classification feature set was the node-splitting approach.
- (4) All the decision trees that were generated were combined to establish a random forest algorithm. The final result was determined through a voting process based on the outcomes of each tree, with the classification result being the one that garnered the majority of votes. A key feature of random forests is their ability to provide relative weights for susceptibility assessment factors, calculated using the Gini index. This index serves as a measure of optimal segmentation in the classification tree, and the importance of factor k is determined by the reduction in the Gini index at node splitting, DG_k . The average importance of factor k is calculated across all trees, summing DG_k for all nodes in the forest. The importance of each evaluation factor is then measured by the percentage of its average Gini reduction value to the sum of all factors' average Gini reduction values, as calculated in Equation (3):

$$P_K = \frac{\sum_{h=1}^n \sum_{j=1}^t D_{Gkhj}}{\sum_{k=1}^m \sum_{h=1}^n \sum_{j=1}^t D_{Gkhj}} \quad (3)$$

where m , n , and t represent the total number of evaluation factors, the number of classification trees, and the number of nodes in a single tree, respectively; D_{Gkhj} denotes the reduced Gini index value of the k -th evaluation factor at the j -th node of the h -th tree; and P_k signifies the importance of the k -th evaluation factor among all evaluation factors.

3.1.3. Weighted Linear Combination

The linear weighted combination method fundamentally utilizes a linear model for comprehensive evaluation. Its simplicity and ease of understanding, coupled with its effective integration with GIS technology, account for its widespread application in this domain. Consequently, this paper adopts the method of information volume model combination for weighting. The formula is as follows:

$$y = \sum_{j=1}^m P_j x_j \quad (4)$$

where y is the value of the integrated information quantity; m refers to the number of evaluation factors; and P_j indicates the random forest weight coefficient corresponding to the evaluation factor x_j , calculated using Equation (3).

3.2. Data Sources

The sources and formats of the evaluation factors used in this study are shown in Table 1.

Table 1. Data sources for the evaluation indicators.

Data Name	Data Source
Elevation/(m)	Geospatial Data Cloud Download DEM
Slope/(°)	Geospatial Data Cloud Download DEM
Aspect/(°)	Geospatial Data Cloud Download DEM
Road network	National Geographic Information Resource Catalog System Download Vector Data
River Distribution	National Geographic Information Resource Catalog System Download Vector Data
NDVI	Geospatial Data Cloud Download Landsat8 Remote Sensing Imagery
Stratigraphic lithology	1:200,000 geological map
Tectonic distribution	1:200,000 geological map

4. Evaluation of Geological Disaster Susceptibility

4.1. Evaluation Unit

Currently, raster units, slope units, and administrative units are used for vulnerability evaluation. In this study area, based on the actual conditions of the area and previous studies of slope-type hazards, we chose the raster unit as the evaluation and analysis unit. Raster unit division is simple, objective, and highly accurate. To maintain the consistency of the evaluation units in the study, the image element size can be calculated by the following formula:

$$G_S = 7.49 + 0.0006S - 2.0 \times 10^{-9}S^2 + 2.9 \times 10^{-15}S^3 \quad (5)$$

where G_S denotes the empirical raster size and S denotes the denominator of the scale in the topographic map used to generate the DEM. The calculations revealed the selection of a 30×30 m raster cell.

4.2. Selection of Evaluation Factors

The precision of geological disaster evaluation hinges on the careful selection of evaluation factors and models. Therefore, to ensure the accuracy of our assessment, it is imperative to choose appropriate evaluation factors and models based on authentic and dependable data. In this paper, previous research and field investigation data were combined to determine the disaster distribution and disaster mechanism, and the geographic environment, geological conditions and human engineering activities in the area, elevation, slope, aspect, NDVI, stratum lithology, distance from the road, distance from the river, and distance from fault were considered as the evaluation indices. The factor grading is shown in Figure 2.

(1) Elevation: The controlling effect of elevation on geological hazards is mainly manifested by the fact that, on the one hand, elevation affects the distribution of groundwater, especially the distribution of underlying water layers, and the groundwater in slopes composed of loose rock and soil bodies is mostly phreatic water; additionally, the higher the elevation is, the less phreatic water is distributed, and the lower the impact on the slopes. On the other hand, elevation plays a controlling role in the range of human activities, and humans mostly live in lower elevation areas and carry out production activities; these factors affect the development of geological hazards. Based on the spatial distribution pattern of hazard points and elevations, the elevation is categorized as <600 m, 600–900 m, 900–1200 m, 1200–1500 m, 1500–1800 m, or >1800 m (Figure 2a).

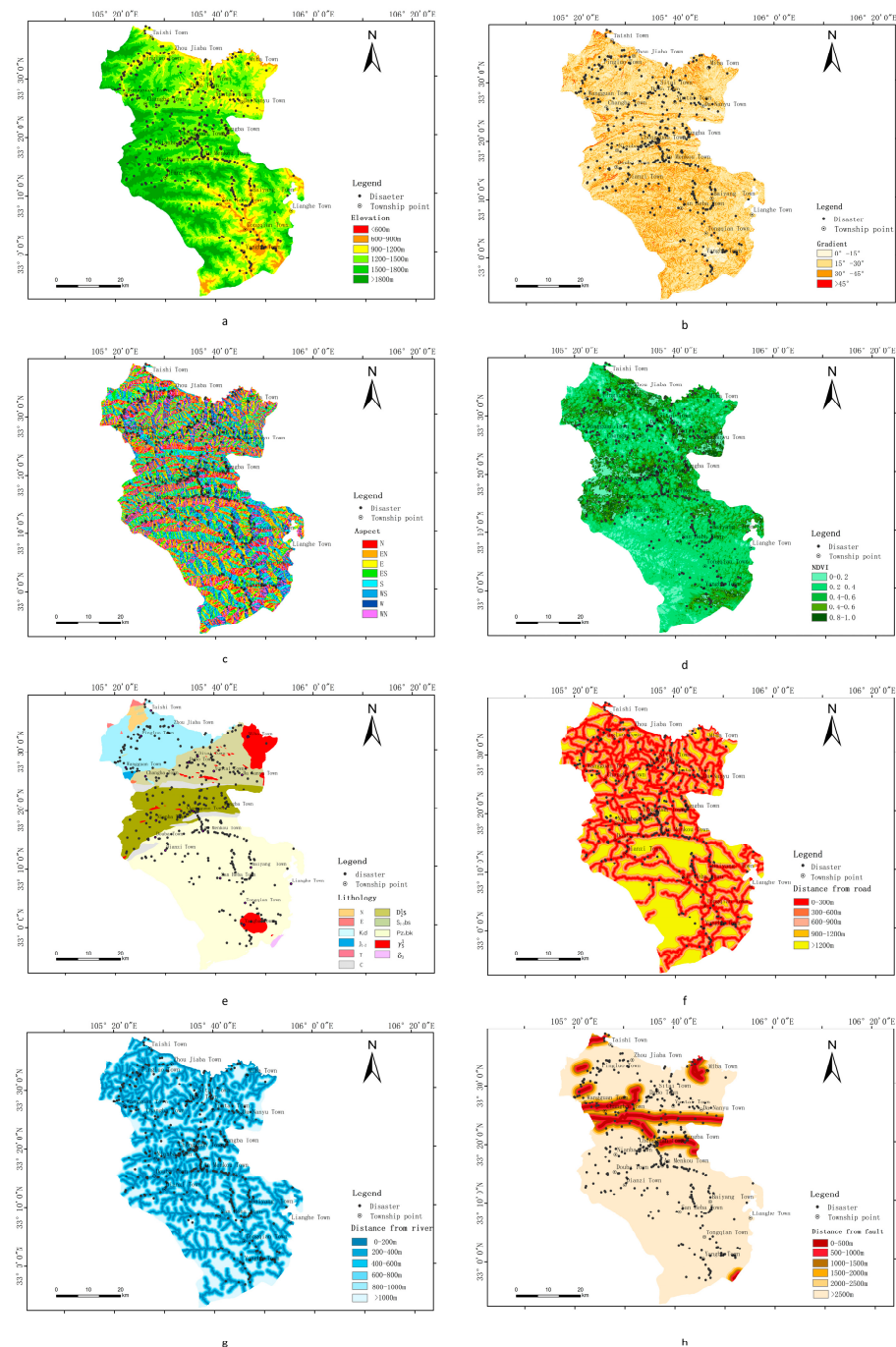


Figure 2. Evaluation factor grading chart.

(2) The slope: The slope gradient is closely related to the occurrence of slope-type geological hazards, and its effect on the shear stress within the geotechnical body is significant. The slope of the study area is divided into 0–15°, 15–30°, 30–45°, and >45° (Figure 2b).

(3) Aspect: Aspect affects light and rainfall, and sunny slopes experience more light and rainfall; thus, the water content of the rock and soil bodies is greater, which reduces the initial conditions for disasters to occur and increases the likelihood of disasters occurring. The aspects of the study area were divided into N, NE, E, SE, S, SW, W, and NW (Figure 2c).

(4) The NDVI plays a role in protecting slopes and preventing soil erosion and has a certain influence on the evolution and stability of slopes; in general, areas with lower vegetation cover are more vulnerable to serious geological hazards, so the normalized difference vegetation index (NDVI) was selected as the evaluation factor (Figure 2d).

(5) **Lithology:** The nature of the rock group, which is the fundamental constituent of the slope, significantly influences the slope stability. The physical and chemical properties of rock bodies often differ for different stratigraphic lithologies, which leads to differences in weathering resistance, erosion capacity, and mechanical properties, thus affecting the occurrence of geological hazards. The study area stratigraphic rock groups are classified as Quaternary (Q), Neoproterozoic (N), Cretaceous (K1d), Jurassic (J1–2), Triassic (T), Carboniferous (C), Devonian (D₂¹S), Silurian (S2–3bs), melanoclastic granite (γ_3^1), pre-Silurian Bikou Group (Pz1bk), and amphibolite (δ_3) (Figure 2e).

(6) **Distance from road:** The areas with frequent human engineering activities in the study area are mostly geological disaster-prone areas, and road construction has damaged the surrounding rock structure and reduced the stability of the surrounding rock. The surface structure was disturbed during road excavation, resulting in structural loosening and increased fissures, which increased the permeability of the slope and increased the risk of disasters. Thus, the distance from roads was included as an evaluation factor, serving as an indicator of the intensity of human engineering activities (Figure 2f).

(7) **Distance from the river:** The study area is located in a mountainous area. The river water volume changes with turbulence, and the hydrodynamic characteristics are obvious. The river scouring effect on the bank embankment easily causes the critical surface to change the forces on the rock and soil body, and the change in the water level easily causes a change in the water content of the soil, which leads to a disaster. Based on the distribution of geological hazards and the characteristics of the water system, the water system is divided into 200 m steps: 0–200 m, 200–400 m, 400–600 m, 600–800 m, 800–1000 m, and >1000 m (Figure 2g).

(8) **Distance from fault:** Geological tectonic factors play a very obvious role in controlling the development of geological hazard sites, and geological hazards are more common in areas where the regional geological structure is more complex, folds are more intense, and new tectonic movements are more active. Geological tectonics determine the distribution of geomorphological formations. The geological tectonic zone contains broken rocks and experiences severe weathering, which damages the continuity and integrity of slopes; additionally, the tectonic zone is the area with the richest and most active groundwater, which reduces the shear strength of the rock body. Additionally, under the action of tectonic stress, joints and fissures develop in the rock body, which provides conditions for the development of geological hazards. Faults widely develop during tectonic movements, so the distance from faults was selected as the evaluation factor. The fault distance was classified as 0–500 m, 500–1000 m, 1000–1500 m, 1500–2000 m, 2000–2500 m, or >2500 m (Figure 2h).

4.3. Multicollinearity Analysis of Evaluation Factors

To understand the relationships between variables and detect multicollinearity, we conducted a multicollinearity analysis. The presence of covariance among variables can distort the differences in the regressors, potentially misrepresenting the relevant factors in the model and, consequently, affecting the evaluation results [25]. Two widely used indicators in the study of multicollinearity are the variance inflation factor (VIF) and the tolerance of variance (TOL).

The VIF quantifies the ratio of the variance of variables in the presence of multicollinearity to the variance of variables in the absence of multicollinearity. This reflects the degree to which the variance increase was caused by multicollinearity. The relationship between TOL and VIF is reciprocal, and a VIF > 10 (i.e., TOL < 0.1) suggests the presence of severe multicollinearity among the variables.

The results of the multicollinearity analysis are presented in Table 2. The maximum VIF is 2.159, while the minimum TOL is 0.463. These findings indicate that there is no significant multicollinearity among the independent variables. Therefore, these variables are considered independent and suitable for use in model applications.

Table 2. Indicators of the multicollinearity test.

Factor	Distance from Road	Distance from River	Lithology	Distance from Fault	Elevation	Slope	Aspect	NDVI
VIF	2.159	2.018	1.187	1.082	1.999	1.064	1.014	1.038
TOL	0.463	0.495	0.843	0.924	0.500	0.940	0.986	0.963

4.4. Calculation of the Value of Information

The data from each evaluation factor were imported into the ArcGIS platform. Some vector data were converted into raster data, which were subsequently classified. The graded raster map of each evaluation factor was superimposed on the disaster point distribution map. The statistical results for each category in each evaluation factor were substituted into Equation (1) to calculate the information contributed by each grade in each evaluation factor to the occurrence of geological hazards. The results are presented in Table 3.

Table 3. Results of the information value.

Evaluation Factor	Level	Information Value	Evaluation Factor	Level	Information Value
Elevation	<600 m	0	Slope	0–15°	0.378954
	600–900 m	1.321535		15–30°	−0.079289
	900–1200 m	0.622475		30–45°	−0.194759
	1200–1500 m	0.009545		>45°	−0.130326
	1500–1800 m	−1.151683		Q	0
Aspect	>1800 m	−2.561344	lithology	N	0
	N	−0.126940		K1d	0.140190
	NE	−0.149518		J1–2	0.179624
	E	−0.197422		T	0.648636
	SE	0.216599		C	−0.7606224
	S	0.217120		D ₂ S	0.562982
	SW	0.021948		S2–3bs	0.297374
	W	0.131493		γ_5^1	−0.660872
	NW	0.224483		δ_3	0
	0–0.2	−0.418219		Pz1bk	−0.278597
NDVI	0.2–0.4	−0.038324	Distance from road	0–300 m	0.953783
	0.4–0.6	0.202757		300–600 m	−0.589427
	0.6–0.8	−0.067471		600–900 m	−1.553367
	0.8–1.0	0.203490		900–1200 m	−1.252783
	0–200 m	0.964993		>1200 m	−1.814372
Distance from river	200–400 m	0.332693	Distance from fault	0–500 m	2.860809
	400–600 m	−1.178137		500–1000 m	0.030943
	600–800 m	−1.662942		1000–1500 m	−0.089632
	800–1000 m	−1.266541		1500–2000 m	0.623365
	>1000 m	−1.442733		2000–2500 m	0.379438
				>2500 m	−2.602948

4.5. Parameter Optimization of the Random Forest Model

(1) Feature Number Selection

Achieving optimal model performance, with the most ideal prediction effect, is contingent upon selecting the optimal parameters. Therefore, determining the ideal number of features is crucial for constructing a random forest model. This selection process relies primarily on the calculated out-of-bag error (OOB error). In this study, the Python programming language was used to iteratively determine the OOB error for different numbers of features in a cyclical manner. The model with the lowest OOB error signifies the highest prediction accuracy. As illustrated in Figure 3, the OOB error is minimized when the number of random features is 3. Consequently, the optimal number of features for this model is identified as three.

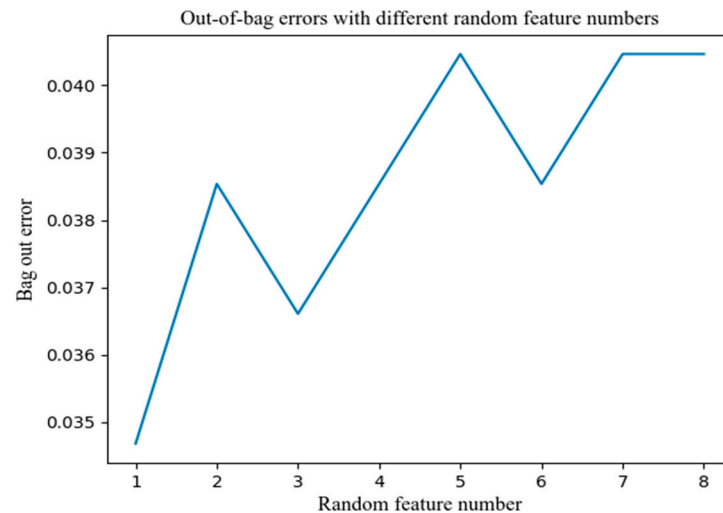


Figure 3. Out-of-bag errors under different characteristic numbers.

(2) Number of Decision Trees

The out-of-bag (OOB) error serves as an unbiased estimator of the generalization error in a random forest. It is calculated as the ratio of the total misclassifications to the total number of samples, representing the OOB error of the random forest. A lower OOB indicates higher model classification accuracy. Python is employed to iterate and calculate the OOB error for varying numbers of decision trees, facilitating the selection of an optimal count. Initially, when the decision tree count is low, the OOB error is high. However, as the number of decision trees increases, the fluctuation range of the OOB error gradually diminishes. Upon reaching a certain number, the OOB error stabilizes, oscillating within a specific range. In this study, the optimal number of decision trees is determined to be 650. The confusion matrix records discrepancies between the actual and predicted model outcomes to assess classification accuracy. As depicted in Figure 4 and Table 4, the model's error rate is 4%, and its accuracy is 95%, calculated as the ratio of correctly classified samples to the total number of samples.

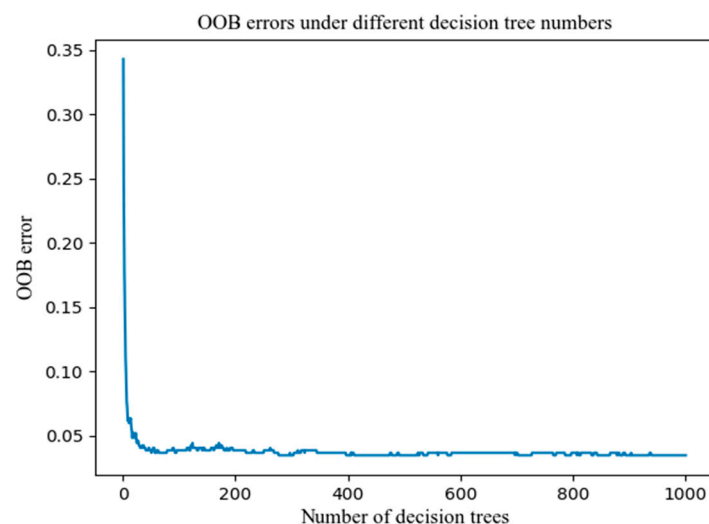


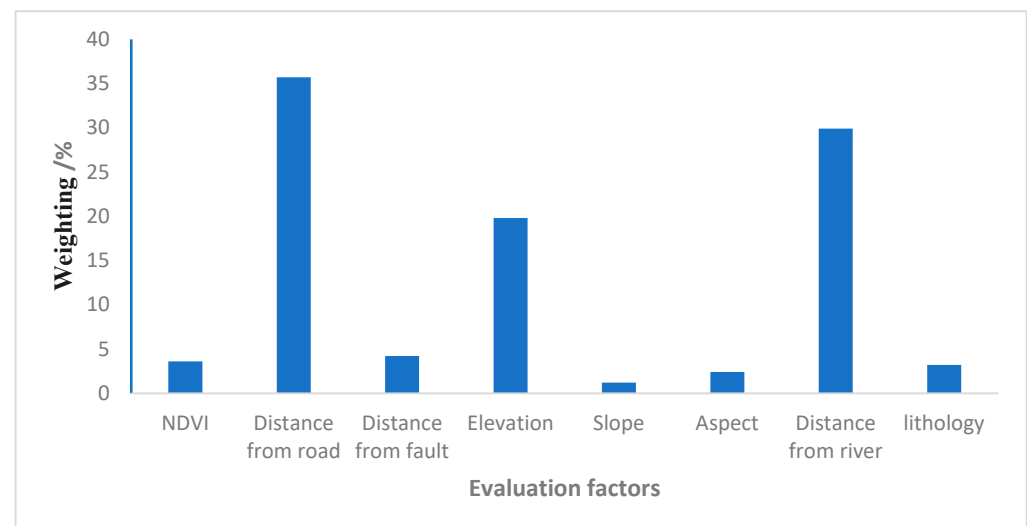
Figure 4. Model error versus number of decision trees.

Table 4. Confusion matrix.

Quantity/Piece		Forecast	
		0	1
Actual	0	99	4
	1	6	114

4.6. Random Forest Weighting

The study area contained 371 disaster points, and 371 non-disaster points were randomly generated based on the results of the informativeness model at a ratio of 1:1, for a total of 742 sample data points. The disaster points are set as “1”, and the non-disaster points are set as “0”. We randomly selected 70% of the sample points as training data and 30% as test data, including disaster points and non-disaster points; subsequently, we substituted them into the random forest model based on PyCharm software (PyCharm 2023.2.1). We calculated the Gini index reduction value of the evaluation factors in the node division and determined the importance of the evaluation factors to determine the factor weights by using Formula (3), thus obtaining the weight diagram of the evaluation factors, as shown in Figure 5.

**Figure 5.** Evaluation factor weighting chart.

4.7. Information Value Weighting

Weight values were calculated by optimizing the random forest algorithm. The weight coefficients were substituted into Equation (4) and linearly combined with the informativeness layer of each evaluation factor to complete the informativeness weighting.

4.8. Geological Hazard Susceptibility Zoning

The study area was divided into low, medium, high, and very high susceptibility zones using the natural discontinuity point method. The geohazard susceptibility zoning map based on the informativeness model and the random forest weighted informativeness model were obtained by regrading the direct and weighted superimposed raster maps according to their informativeness values. Figures 6 and 7 show that the distribution of disaster points aligns well with the susceptibility zones.

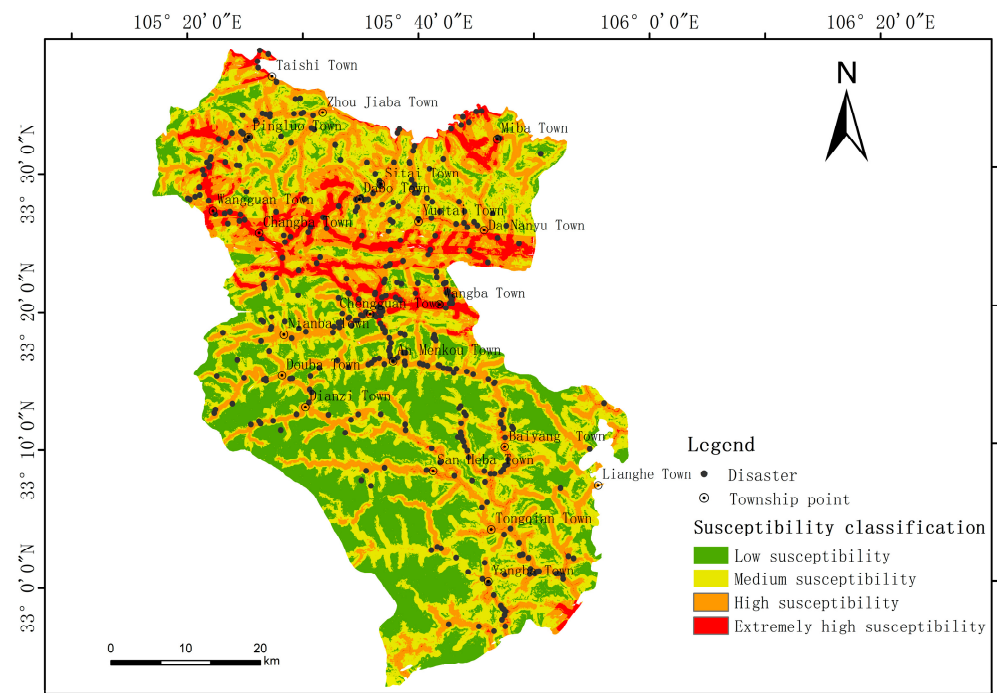


Figure 6. Geological hazard susceptibility zoning map based on the IM.

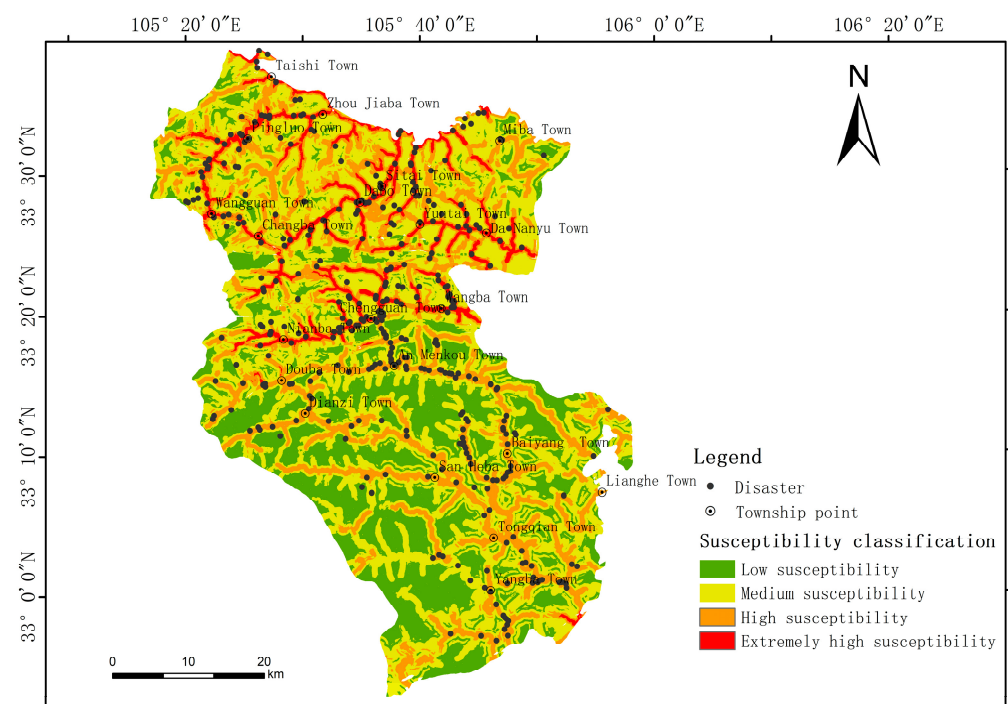


Figure 7. Geological hazard susceptibility zoning map based on RF-weighted IM.

5. Evaluation Findings and Analyses

5.1. Comparative Analysis of Susceptibility Zoning Maps

The two models exhibit similar vulnerability assessment results. In Kang County, areas with high and extremely high susceptibility to geological hazards are primarily located in the northern region and extend in a strip-like formation. The central area, near Kang County's urban zone, is susceptible to geohazards due to urbanization and increased human engineering activities. The southern valley region, characterized by a dense network of valleys and steep terrain, is also prone to geological hazards. Areas along main roads

and rivers, where human engineering activities and riverbank scouring are prevalent, are highly susceptible to disasters. Conversely, low to medium susceptibility areas are mainly in Kangxian County's southern and northern mountainous regions, where vegetation cover is abundant, geological conditions are favorable, and human engineering activities are minimal.

The northern part of Kang County has a dense distribution of disaster sites. The weighted information volume model based on the random forest model aligns well with the distribution of extremely high susceptibility zones, especially along the Pingluo and Yanzi Rivers and adjacent to the Baiwang and Kangyang Highways. These areas have fragile geological environments. The single information volume model, on the other hand, distributes extremely high susceptibility zones along an east–west axis, with less clear division and poorer applicability. Thus, the random forest-based weighted informativeness model is more effective and reliable for susceptibility zone classification.

5.2. Comparative Statistical Analysis of Susceptibility Zones

The geohazard susceptibility zoning maps for Kang County were generated using both the information quantity model and the random forest weighted information quantity model. These maps were then overlaid with historical geohazard sites, and the density of hazard sites within each susceptibility zone was subjected to statistical analysis (Tables 5 and 6). The tables presenting the density of disaster points reveal that as the degree of disaster susceptibility increases, the density of disaster points within the susceptibility zones, as determined by both models, gradually increases. The highest density of disaster points is observed in the extremely high susceptibility zones. These results align with the distribution of actual disaster sites, indicating a favorable evaluation outcome.

Table 5. Table of statistical information on the geological hazard susceptibility zones of IM.

Susceptibility Division	Area/km ²	Disaster Sites/pc	Density of Hazardous Sites/(units/km ²)
Low susceptible area	85.21929	7	0.082
Medium susceptible area	103.82031	53	0.510
High susceptible areas	84.63600	225	2.658
Extremely high susceptible areas	20.45664	86	4.204

Table 6. Table of geological hazard susceptibility zones of RF-weighted IM.

Susceptibility Division	Area/km ²	Disaster Sites/pc	Density of Hazardous Sites/(units/km ²)
Low susceptible area	73.40715	4	0.054
Medium susceptible area	132.69357	64	0.482
High susceptible areas	71.95941	217	3.016
Extremely high susceptible areas	16.07211	86	5.351

A comparison of the random forest weighted information quantity model and the single information quantity model reveals that the two models exhibit similar performances in the low and medium susceptibility zones. However, as the degree of susceptibility increases, the RF-weighted information quantity model yields a higher density of hazard points than does the single information quantity model. This difference is particularly noticeable in the high susceptibility zones and extremely high susceptibility zones, where the random forest weighted information quantity model yields hazard point densities that increase by 0.358/km² and 1.147/km², respectively.

In summary, the random forest weighted information quantity model exhibited a denser distribution of hazard points in high vulnerability areas, resulting in evaluation results that more closely aligned with the actual distribution of hazard points (Figure 8).

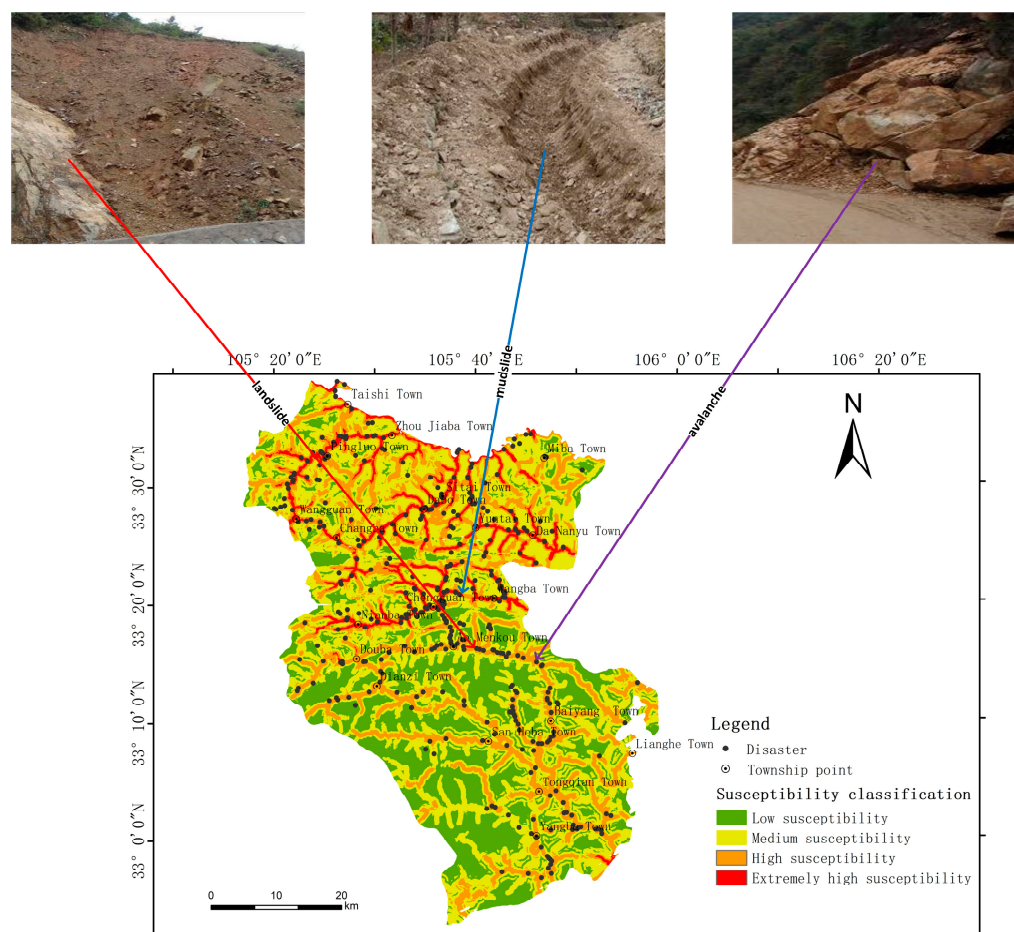


Figure 8. Field validation.

5.3. Accuracy Check Analysis

The ROC curve, originally derived from statistical decision theory, is also referred to as the susceptibility curve or the subject characteristic curve. As an outcome evaluation method, it possesses the advantage of not being constrained by critical limitations. The results accurately depict the relationship between the specificity and sensitivity of the applied analysis method and exhibit commendable experimental precision. Consequently, this approach has extensive utility in geohazard susceptibility assessments [26]. Analyses were conducted with SPSS software (SPSS 26), with the area under the ROC curve (AUC) serving as an indicator reflecting and comparing the predictive accuracy of the models (see Figure 9). The AUC values for the informativeness model and the random forest weighted informativeness model were 0.934 and 0.951, respectively, indicating predictive accuracies of 93.4% and 95.1%, respectively. The analytical findings indicate that the random forest weighted information quantity model outperforms the information quantity model in assessing slope-type geological hazard susceptibility.

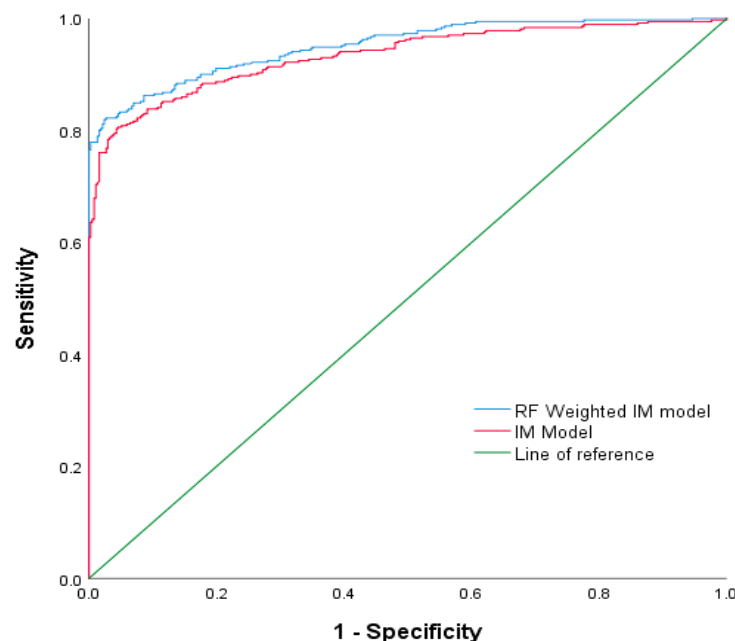


Figure 9. ROC curve.

6. Discussion

Currently, both domestic and international scholars employ various models to assess susceptibility to geological hazards, generally categorized into quantitative and qualitative methods [27–29]. Qualitative methods heavily rely on subjective expertise and individual analytical judgment, resulting in potentially divergent and inconsistent results when different experts are involved. On the other hand, quantitative methods face challenges in determining the interrelationships among factors in high-dimensional space. In this study, we combined the information quantity model with the random forest model using the strengths of both statistical analysis and integrated learning. Our goal is to reduce the influence of subjective factors, clarify the significance of evaluation factors, enhance the accuracy and reliability of disaster evaluation, and ultimately predict susceptibility to geological disasters in the study area.

In the process of conducting slope-type geological hazard susceptibility assessments using the random forest model, selecting non-hazardous points is challenging because determining the specific locations of non-hazardous areas is not straightforward. Previous studies often selected non-hazard points by random sampling in the entire area or by creating buffers around hazard points. However, these approaches did not consistently yield optimal non-hazardous points. To improve the model accuracy, we utilized the information quantity model to quickly evaluate the susceptibility of the study area and selected non-hazardous points within the low-level susceptibility category. This approach allowed us to obtain non-hazardous points with a “high probability”, which were subsequently applied in the random forest weighted information quantity model. This approach significantly reduces the risk of misclassifying potential hazardous points as non-hazardous.

In this study, we determined the optimal parameters for the study area through iterative processes and calculated the weights of the evaluation factors to enhance the accuracy of the evaluation results. This paper has several limitations.

The accuracy of the model is well demonstrated in this study area, which is theoretically generalizable; however, no other study area has been selected to verify the generalizability of the model. Multiple study areas were selected for validation.

Although the weights of each evaluation factor were accurately calculated in the present study, it is still impossible to exclude the influence of human subjective factors during factor selection, which can be solved at a later stage by analyzing the geological

conditions of the breeding disaster, the mechanism of the disaster, and the mode of the disaster.

The complex topographic conditions of the study area and the difficulties in obtaining certain geological information pose challenges that need to be considered.

7. Conclusions

The evaluation of geological hazard susceptibility is crucial for implementing timely protective measures to reduce potential damage. In this study, we selected eight evaluation factors based on previous research findings and the characteristics of disaster development. Utilizing a GIS platform, we employed the information quantity model and the random forest weighted information quantity model to assess susceptibility to slope-type geological hazards in the study area, resulting in the following conclusions:

- (1) Both the information quantity model and the random forest weighted information quantity model exhibit an increase in the density of hazard points within susceptibility zones as hazard susceptibility increases. The highest density of hazard points is observed in the extremely high susceptibility zone. Compared to the single information quantity model, the weighted information quantity model yields a greater density of hazard points in high susceptibility zones. According to the ROC curve, both models achieved reasonable results, with AUC values of 0.934 and 0.951, respectively, meeting the test requirements. However, the latter exhibited higher ROC accuracy, indicating that the weighted informativeness model aligns more closely with the distribution of historical disaster points in the susceptibility assessment results, resulting in better delineation of susceptibility zones.
- (2) The areas classified as having high or extremely high susceptibility to geological disasters in Kang County are primarily located in the northern part of the county and areas with frequent human engineering activities; these areas account for 35.73% of the total study area. These regions are at high risk of geological disasters and warrant close attention from relevant authorities. The geological disaster susceptibility zoning map of the study area allows for the prediction of disaster likelihood in different areas, offering valuable insights for policymaking, disaster prevention, relief deployment, and property loss reduction, ultimately safeguarding lives. Additionally, this study provides essential guidance for local urban development and spatial planning and promotes sustainable social and regional development.
- (3) The random forest model is easy to use, robust against outliers and noise, and has high predictive accuracy and stability. Moreover, this approach is insensitive to multicollinearity, thus mitigating the risk of overfitting. Importance scores are provided for each evaluation factor, with road buffer distance, river buffer distance, and elevation emerging as significant factors for disaster occurrence in this study. Geological hazards in the study area are predominantly manifested within a 900 m proximity to the road. This is attributed to frequent human engineering activities in this region, which have resulted in numerous artificial cut slopes for road construction and house building, thereby exacerbating landslide occurrences. Closer proximity to the river increases disaster susceptibility due to erosion, scouring effects, and groundwater softening on the rock and soil bodies. Within the elevation range of 900–1500 m, frequent human engineering activities, driven by the density of villages and towns, has led to numerous bare soil slopes, further exacerbating disaster occurrences.
- (4) Based on the information volume model's partitioning results, areas with high and extremely high susceptibility are excluded. Non-hazardous points are randomly selected from the low and medium susceptibility areas. This selection serves the dual purpose of reducing the likelihood of misclassified potential hazard points as non-hazardous and preventing overfitting of the random forest model, which can occur when selecting non-disaster points in low susceptibility areas. Consequently, this approach enhances the model's predictive accuracy. Optimal feature selection is determined by calculating the out-of-bag error under different feature numbers, while

the number of decision trees is chosen iteratively through out-of-bag error calculations under varying tree quantities. Model parameters are evaluated in conjunction with the confusion matrix, leading to the identification of optimal RF parameters. This process aims to reduce overfitting, enhance the model's generalization capacity, adapt the model to the study area, and improve the overall model efficiency.

- (5) The evaluation model developed in this study can be extended to other areas with geohazard potential. Prior to conducting evaluations, comprehensive data on geological structure, stratigraphic lithology, meteorology, hydrology, and human engineering activities must be collected. Adjustments to individual evaluation factors and model parameters based on the study area's characteristics are essential. Comparing historical disaster sites with evaluation results helps evaluate the model's prediction effectiveness. If substantial disparities exist, further adjustments and improvements are necessary. It is important to recognize that different regions have unique geological, environmental, and disaster development characteristics, leading to variations in model predictions. These regional differences should be considered, and the model should be adapted and refined to optimize its performance in specific areas.

Author Contributions: Material preparation, data collection, and analysis were performed by S.T., R.L., M.Z., S.Z., L.Z. and H.W. The first draft of the manuscript was written by R.L. and all authors commented on previous versions of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Science and Technology Innovation Team Program of Yunnan Province Education Department (Grant No. CY22624109); the Graduate Tutor Team Program of Yunnan Province Education Department (Grant No. CY22622205); the First Class University Program of Yunnan University (Grant No. CY22624108); and the Xing Dian Talent Teacher's Program of Yunnan Province (Grant No. XDTT202206).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The DEM and NDVI data are openly available in the geospatial data cloud platform at <https://www.gscloud.cn/> (accessed on 2 October 2023). Geological hazard point data and 1:200,000 geological maps are openly available from the Geographic Remote Sensing Ecology Network platform at <http://www.gisrs.cn/> (accessed on 2 October 2023). Water system and road data can be obtained from the National Catalogue Service for Geographic Information at <https://www.webmap.cn/> (accessed on 3 October 2023).

Acknowledgments: We would like to thank laboratory colleagues for their hard work and our mentor for their careful guidance. We would also like to express our gratitude for the funding we received.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Xu, Q.; Huang, R.; Wang, L. Mechanistic analysis of geological disasters induced by external disturbance. *J. Rock Mech. Eng.* **2002**, *21*, 280–284. [\[CrossRef\]](#)
- Galvani, A.P.; Bauch, C.T.; Anand, M.; Singer, B.H.; Levin, S.A. Human-environment interactions in population and ecosystem health. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 14502–14506. [\[CrossRef\]](#) [\[PubMed\]](#)
- Wang, N.; Wang, Y.; Luo, D.; Yao, Y. A review of research on landslide prediction and forecasting in China. *Geol. Rev.* **2008**, *54*, 355–361. [\[CrossRef\]](#)
- Lara, M.; Sepúlveda, S.A. Landslide susceptibility and hazard assessment in San Ramón Ravine, Santiago de Chile, from an engineering geological approach. *Environ. Earth Sci.* **2010**, *60*, 1227–1243. [\[CrossRef\]](#)
- Lucà, F.; D'Ambrosio, D.; Robustelli, G.; Rongo, R.; Spataro, W. Integrating geomorphology, statistic and numerical simulations for landslide invasion hazard scenarios mapping: An example in the Sorrento Peninsula (Italy). *Comput. Geosci.* **2014**, *67*, 163–172. [\[CrossRef\]](#)
- Calvello, M.; d'Orsi, R.N.; Piciullo, L.; Paes, N.; Magalhaes, M.; Lacerda, W.A. The Rio de Janeiro early warning system for rainfall-induced landslides: Analysis of performance for the years 2010–2013. *Int. J. Disast. Risk Reduct.* **2015**, *12*, 3–15. [\[CrossRef\]](#)
- Salciarini, D.; Godt, J.W.; Savage, W.Z.; Conversini, P.; Baum, R.L.; Michael, J.A. Modeling regional initiation of rainfall-induced shallow landslides in the eastern Umbria Region of central Italy. *Landslides* **2006**, *3*, 181–194. [\[CrossRef\]](#)

8. Patriche, C.V.; Pirnau, R.; Grozavu, A.; Rosca, B. A comparative analysis of binary logistic regression and analytical hierarchy process for landslide susceptibility assessment in the Dobrov River Basin, Romania. *Pedosphere* **2016**, *26*, 335–350. [\[CrossRef\]](#)
9. Chen, W.; Xie, X.; Peng, J.; Wang, J.; Duan, Z.; Hong, H. GIS-based landslide susceptibility modelling: A comparative assessment of kernel logistic regression, Naïve-Bayes tree, and alternating decision tree models. *Geomat. Nat. Hazards Risk* **2017**, *8*, 950–973. [\[CrossRef\]](#)
10. Ozdemir, A. A comparative study of the frequency ratio, analytical hierarchy process, artificial neural networks and fuzzy logic methods for landslide susceptibility mapping: Taşkent (Konya), Turkey. *Geotech. Geol. Eng.* **2020**, *38*, 4129–4157. [\[CrossRef\]](#)
11. Aditian, A.; Kubota, T.; Shinohara, Y. Comparison of GIS-based landslide susceptibility models using frequency ratio, logistic regression, and artificial neural network in a tertiary region of Ambon, Indonesia. *Geomorphology* **2018**, *318*, 101–111. [\[CrossRef\]](#)
12. Pradhan, B. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. *Comput. Geosci.* **2013**, *51*, 350–365. [\[CrossRef\]](#)
13. Dou, J.; Yamagishi, H.; Pourghasemi, H.R.; Yunus, A.P.; Song, X.; Xu, Y.; Zhu, Z. An integrated artificial neural network model for the landslide susceptibility assessment of Osado Island, Japan. *Nat. Hazards* **2015**, *78*, 1749–1776. [\[CrossRef\]](#)
14. Su, C.; Wang, L.; Wang, X.; Huang, Z.; Zhang, X. Mapping of rainfall-induced landslide susceptibility in Wencheng, China, using support vector machine. *Nat. Hazards* **2015**, *76*, 1759–1779. [\[CrossRef\]](#)
15. Liu, R.; Shi, S.; Sun, D.; Xu, J. Landslide susceptibility zoning in Wushan County based on GIS and random forest. *J. Chongqing Norm. Univ. (Nat. Sci. Ed.)* **2020**, *37*, 86–96. [\[CrossRef\]](#)
16. Ahmed, N.; Firoze, A.; Rahman, R.M. Machine learning for predicting landslide risk of Rohingya refugee camp infrastructure. *J. Inform. Telecommun.* **2020**, *4*, 175–198. [\[CrossRef\]](#)
17. Merghadi, A.; Abderrahmane, B.; Tien Bui, D. Landslide susceptibility assessment at Mila Basin (Algeria): A comparative assessment of prediction capability of advanced machine learning methods. *ISPRS Int. J. Geo-Inform.* **2018**, *7*, 268. [\[CrossRef\]](#)
18. Zheng, Y.; Chen, J.; Wang, C.; Cheng, T. Application of coefficient of determination and random forest model in landslide susceptibility assessment in Mangshi, Yunnan. *Geosci. Technol. Bull.* **2020**, *39*, 131–144. [\[CrossRef\]](#)
19. Lin, R.; Liu, J.; Xu, S.; Liu, M.; Zhang, M.; Liang, E. Landslide susceptibility evaluation method based on stochastic forest weighted information. *Surv. Mapp. Sci.* **2019**, *45*, 131–138. [\[CrossRef\]](#)
20. Liu, Y.; Di, B.; Zhan, Y.; Stamatoopoulos, C.A. Evaluation of mudslide susceptibility based on random forest model: An example of Wenchuan earthquake-hit area. *J. Mt. Geogr.* **2018**, *36*, 765–773. [\[CrossRef\]](#)
21. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
22. Goetz, J.N.; Brenning, A.; Petschko, H.; Leopold, P. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Comput. Geosci.* **2015**, *81*, 1–11. [\[CrossRef\]](#)
23. Dou, J.; Yunus, A.P.; Bui, D.T.; Merghadi, A.; Sahana, M.; Zhu, Z.; Chen, C.W.; Khosravi, K.; Yang, Y.; Pham, B.T. Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan. *Sci. Total Environ.* **2019**, *662*, 332–346. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Verikas, A.; Gelzinis, A.; Bacauskiene, M. Mining data with random forests: A survey and results of new tests. *Pattern Recogn.* **2011**, *44*, 330–349. [\[CrossRef\]](#)
25. Hu, T.; Fan, X.; Wang, S.; Guo, Z.; Liu, A.; Huang, F. Evaluation of landslide susceptibility in Sinan County based on logistic regression model and 3S technolog. *Geosci. Technol. Bull.* **2022**, *39*, 113–121. [\[CrossRef\]](#)
26. Sun, D. Research on Landslide Susceptibility Zoning and Rainfall-Induced Landslide Forecasting and Warning Based on Machine Learning. Ph.D. Thesis, East China Normal University, Shanghai, China, 2019.
27. Du, Q.; Fan, W.; Li, K.; Yang, D.; Lv, J. Application of binary logistic regression and informativeness model in geological hazard zoning. *Disaster Sci.* **2017**, *32*, 220–226. [\[CrossRef\]](#)
28. Pawluszek, K.; Borkowski, A. Impact of DEM-derived factors and analytical hierarchy process on landslide susceptibility mapping in the region of Rożnów Lake, Poland. *Nat. Hazards* **2017**, *86*, 919–952. [\[CrossRef\]](#)
29. Hong, H.; Pradhan, B.; Sameen, M.I.; Kalantar, B.; Zhu, A.; Chen, W. Improving the accuracy of landslide susceptibility model using a novel region-partitioning approach. *Landslides* **2018**, *15*, 753–772. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.