

Influences of non-landslide sample selection strategies on landslide susceptibility mapping by machine learning

Hongqiang Dou, Junbin He, Siyi Huang, Wenbin Jian & Chaoxu Guo

To cite this article: Hongqiang Dou, Junbin He, Siyi Huang, Wenbin Jian & Chaoxu Guo (2023) Influences of non-landslide sample selection strategies on landslide susceptibility mapping by machine learning, *Geomatics, Natural Hazards and Risk*, 14:1, 2285719, DOI: [10.1080/19475705.2023.2285719](https://doi.org/10.1080/19475705.2023.2285719)

To link to this article: <https://doi.org/10.1080/19475705.2023.2285719>



© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 27 Nov 2023.



[Submit your article to this journal](#)



Article views: 2874



[View related articles](#)



[View Crossmark data](#)



Citing articles: 45 [View citing articles](#)



Influences of non-landslide sample selection strategies on landslide susceptibility mapping by machine learning

Hongqiang Dou^{a,b}, Junbin He^a, Siyi Huang^a, Wenbin Jian^{a,b} and Chaoxu Guo^b

^aZijin School of Geology and Mining, Fuzhou University, Fuzhou, Fujian, China; ^bFujian Key Laboratory of Geohazard Prevention, Fuzhou, Fujian, China

ABSTRACT

Landslide susceptibility mapping is crucial in mitigating the risk of regional landslide hazards. The main objectives of this study are to design and analyze the impact of three non-landslide data generation methods on landslide susceptibility mapping. The specific non-landslide data selection methods are the random generation of non-landslide points through different distance landslide buffers, by selecting areas with low landslide influence in the landslide condition factor layer, and based on the results of the information value model partitioning by selecting low to medium susceptibility zones. There were 14 landslide condition factors used for the landslide susceptibility mapping, and correlations between landslides and condition factors were analyzed using the Gini index. The 70% dataset was modelled using a logistic regression model as well as an artificial neural network model. Finally, statistical metrics and AUC values were used for the 30% validation data for comparing the predictive performances of different non-landslide data generation strategies in the model. Based on the validation results, models A6 and B6 performed the best, with respective AUC values of 0.997 and 0.995. The findings show that the non-landslide samples generated using the low susceptibility interval of the information value model have the best performance.

ARTICLE HISTORY

Received 20 September 2023
Accepted 16 November 2023

KEYWORDS

Non-landslide sample;
selection method; landslide
susceptibility mapping;
machine learning

1. Introduction

Landslide disasters have claimed many lives and damaged much property around the world. Just in 2020, the landslides in China caused 197 casualties and direct economic damages of more than RMB 5 billion. The coastal areas of southeast China are highly vulnerable to landslide disasters due to typhoons and heavy rainfall, which seriously threaten people's lives and property. Researchers have suggested various solutions to mitigate the adverse consequences of landslides. Landslide susceptibility maps are effective tools for identifying and predicting potential landslides. Research on

CONTACT Hongqiang Dou  douhq@fzu.edu.cn

© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

landslide susceptibility zoning is significant in mitigating the damage to people and property caused by landslides (Zêzere et al. 2017; Huang and Zhao 2018; Azarafza et al. 2021; Yavuz Ozalp et al. 2023).

During the past few decades, several studies have been conducted on landslide susceptibility assessment using various methods, including deterministic coefficients (Chong 2010; Pradhan et al. 2017), linear regression (Onagh et al. 2012), the weight of evidence (Lee and Choi 2004; Ilia and Tsangaratos 2016), analytic hierarchy process (Kayastha et al. 2013), and frequency ratios (Mohammady et al. 2012; Akinci and Yavuz Ozalp 2021). Susceptibility mapping has also used machine learning models such as logistic regression (Umar et al. 2014), support vector machines (Chen et al. 2017; Zhou et al. 2018), random forests (Youssef et al. 2016; Chen et al. 2018), and neural networks (Pradhan and Lee 2010; Xu et al. 2015; Akinci 2022). Besides, Hybrid models developed by combining the two methods have become a hot topic for research on susceptibility evaluation models, including bivariate statistical methods (Constantin et al. 2011; Schicker and Moon 2012), neural network fuzzy logic, EBF fuzzy logic, and fuzzy evidence weighting (Hong et al. 2017; Hong et al. 2018).

The above literature review shows that though the published research has come up with enormous outcomes in landslide susceptibility mapping models, the importance of susceptibility sample selection in model prediction accuracy has not been given much attention. It is necessary to prepare modeling and validation datasets before evaluating the landslide susceptibility. These dataset samples include landslide points, non-landslide points, and condition factors. The sample selection significantly impacts the accuracy and reasonability of the susceptibility zoning (Dou et al. 2020; Gaidzik and Ramírez-Herrera 2021). To solve this challenge, many scholars (Lima et al. 2017; Li et al. 2021; Wang et al. 2022) have conducted experimental studies on the selection strategy of non-landslide point selection to address this problem. The most commonly used methods for selecting non-landslide samples are (1) random sampling method, in which non-landslides other than known landslides in the study area are selected randomly (Pham et al. 2016), (2) buffer sampling method, in which non-landslide points are selected based on known landslide samples, by buffering outward method (Peng et al. 2014; Su et al. 2017) and (3) river or slope sampling method, which considers the selection of non-landslide samples in the rivers or areas with slopes less than 2° (Kavzoglu et al. 2014). The selection of non-landslide samples may emerge in landslide-prone locations due to these methodologies' relatively random and subjective nature, which lowers the accuracy of landslide prediction.

In view of this, this study improves the above three methods, proposes a non-landslide sample selection method based on buffer, condition factor and information value models, and establishes a study area susceptibility model using logistic regression and artificial neural network, respectively. The efficacy of the non-landslide sample selection methods was then evaluated and compared by contrasting the accuracy and logic of distinct susceptibility models using various evaluation methods. Eventually, the most reasonable non-landslide sample selection method is presented.

2. Study area and data source

2.1. Study area

The study area is located in Anxi County, Quanzhou City, Fujian Province, China (Figure 1). Anxi County between $25^{\circ}50'36''$ - $26^{\circ}26'30''$ north latitude and $117^{\circ}48'30''$ - $118^{\circ}40'01''$ east longitude, has a total area of 3057.28 km^2 and is 63 km long from north to south and with an east-west width of 74 km and an altitude ranging from 8 m to 1589 m. The study area is exposed to differences in topographic condition and the strength of marine climatic influences owing to its location in a subtropical monsoon climate zone. The average annual precipitation is above 1000 mm, and the rainy season extends from April to June. Typhoons and tropical storms make landfall about 1 to 3 times/year due to the region's close proximity to the southeast coast. Typhoon-induced heavy-very heavy rainfall is the main triggering factor of landslide disaster outbreaks.

2.2. Data source

The geological and historical landslide data were extracted from the list of landslides gathered from the Anxi County Bureau of Land and Resources. A total of 821

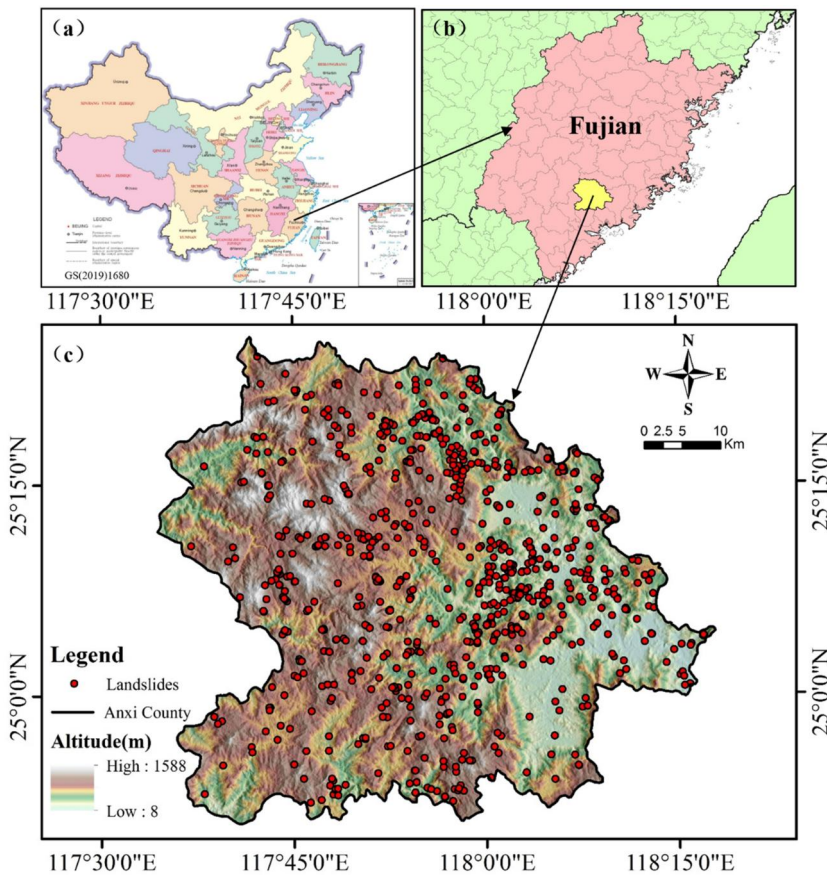


Figure 1. Location of the study area and landslide inventory map: (a) Administrative divisions of China, (b) location of the study, (c) landslide inventory.

historical landslides were identified in the study area (Figure 1(c)), among which the most extensive and smallest landslide volumes were $136,000 \text{ m}^3$ and 12 m^3 , respectively. Most of the monitored landslide locations were close to the front and back of residents' houses for the safety of residents and public facilities.

Various factors should be considered by the evaluation index system of landslide susceptibility in the study area. In addition to topographic factors (elevation, slope angle and curvature) and geological factors (lithology and faults), rainfall characteristics and human engineering operations, such as road distance, are taken into account. Finally, for this study, 14 initial conditional factors for landslide susceptibility are chosen (Figure 2). Specific information on each factor is presented in Table 1.

Slope gradient is one of the most important factors influencing the occurrence and development of landslides and their morphological characteristics due to its considerably varying degrees of influence on surface water runoff, groundwater recharge and discharge, material transport and accumulation, and stress distribution characteristics of the slope body. The slope elevation was derived from the DEM and classified into six categories with a spacing of 10° for this investigation (Figure 2(a)).

The elevation is also vital in inducing landslides due to its control on slope progradation. Landslides are less likely to occur in very low elevations with thick sedimentary soils and in higher elevations with very hard rocks. In contrast, the probability of landslide occurrence is higher in the middle altitude areas, where the slope body is mostly a binary structure of upper soil and lower rocks, with frequent human engineering activities (Figure 2(b)).

Different slope orientations produce different vegetation covers and soil moisture contents due to changes in the time and intensity of solar illumination. The slope orientation is derived from the DEM, which is divided into nine directional classes, namely flat, north, north-east, east, south-east, south, west, south-west and north-west (Figure 2(c)).

Normalized difference vegetation index (NDVI), which represents the vegetation growth status, reflects the sparseness of surface vegetation and is often used to distinguish vegetation cover areas, bare ground, and water bodies. The NDVI of the study area was derived from Landsat8 satellite maps in the near-infrared band and infrared band (Figure 2(d)).

Curvature is the second-order derivative of the surface and is divided into plane curvature and profile curvature, representing the curvatures in the direction parallel to and perpendicular to the maximum slope. The erosion and deposition are impacted by the flow acceleration and deceleration of profile curvature. On the other hand, the planar curvature affects flow convergence and dispersion. The DEM of the study area was used to derive the planar and profile curvatures, which are classified into eight categories based on the natural discontinuity method (Figure 2(e,f)).

Faults cause the development of rock joints and fractures, which results in the deterioration of the geotechnical properties and favors landslide development. In general, the likelihood of landslides increases with the fault's proximity. The 1:50,000 geological map is used to calculate fault distances, which are classified into eight categories with 500 m intervals (Figure 2(g)).

The stratigraphic lithology of an area plays an important role in the formation and development of landslides. It is an intrinsic factor in breeding landslide occurrence, as the variations in lithology impact the degree of landslide development by determining the type and scale characteristics of landslides. The lithology map of Anxi County was created using vectorization in Arcgis based on the 1:50,000 geological

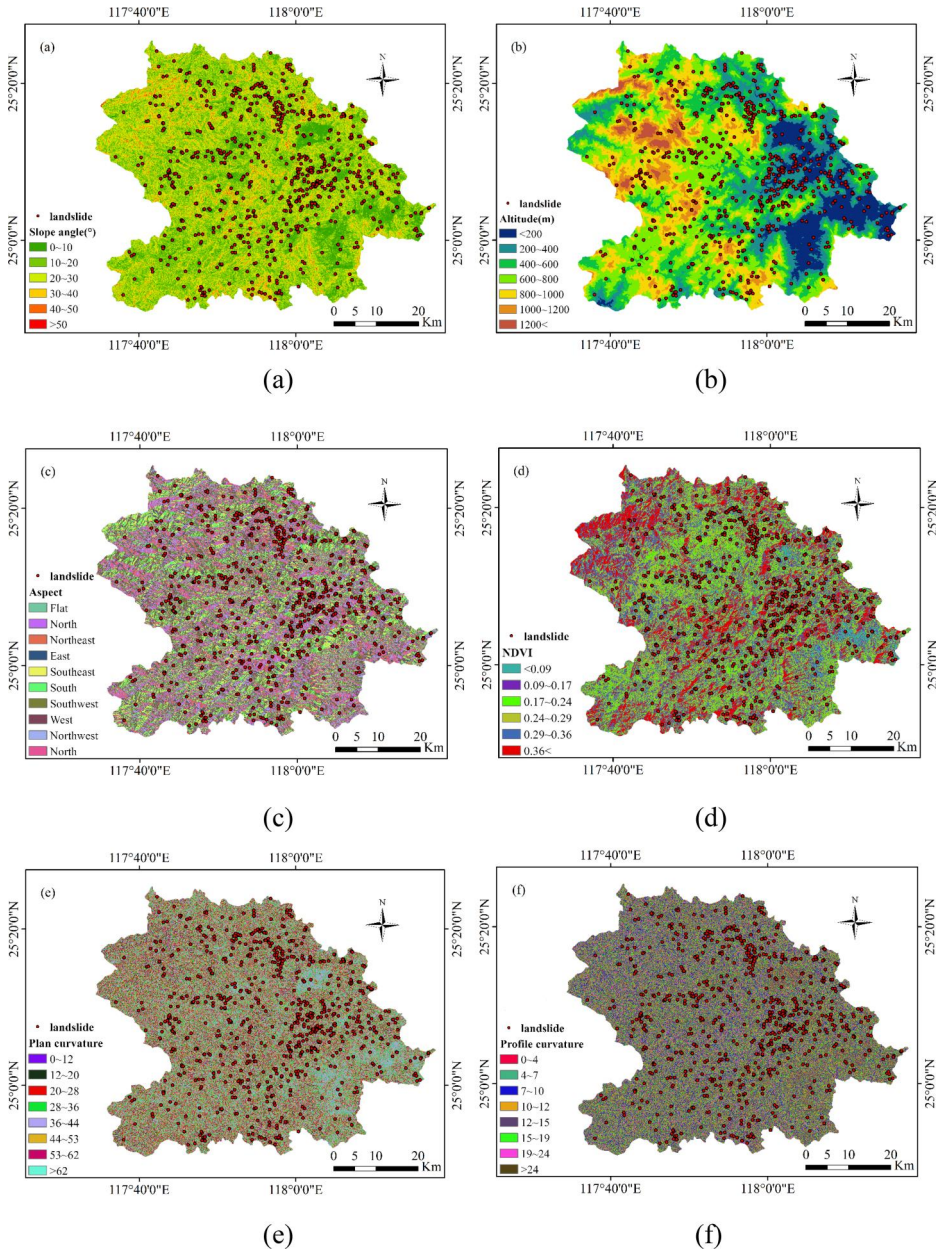


Figure 2. Landslide conditioning factors: (a) slope angle, (b) altitude, (c) aspect, (d) NDVI, (e) plan curvature, (f) profile curvature, (g) distance to faults, (h) lithology, (i) distance to rivers, (j) TWI, (k) rainfall, (l) distance to roads, (m) distance to residences, and (n) land use.

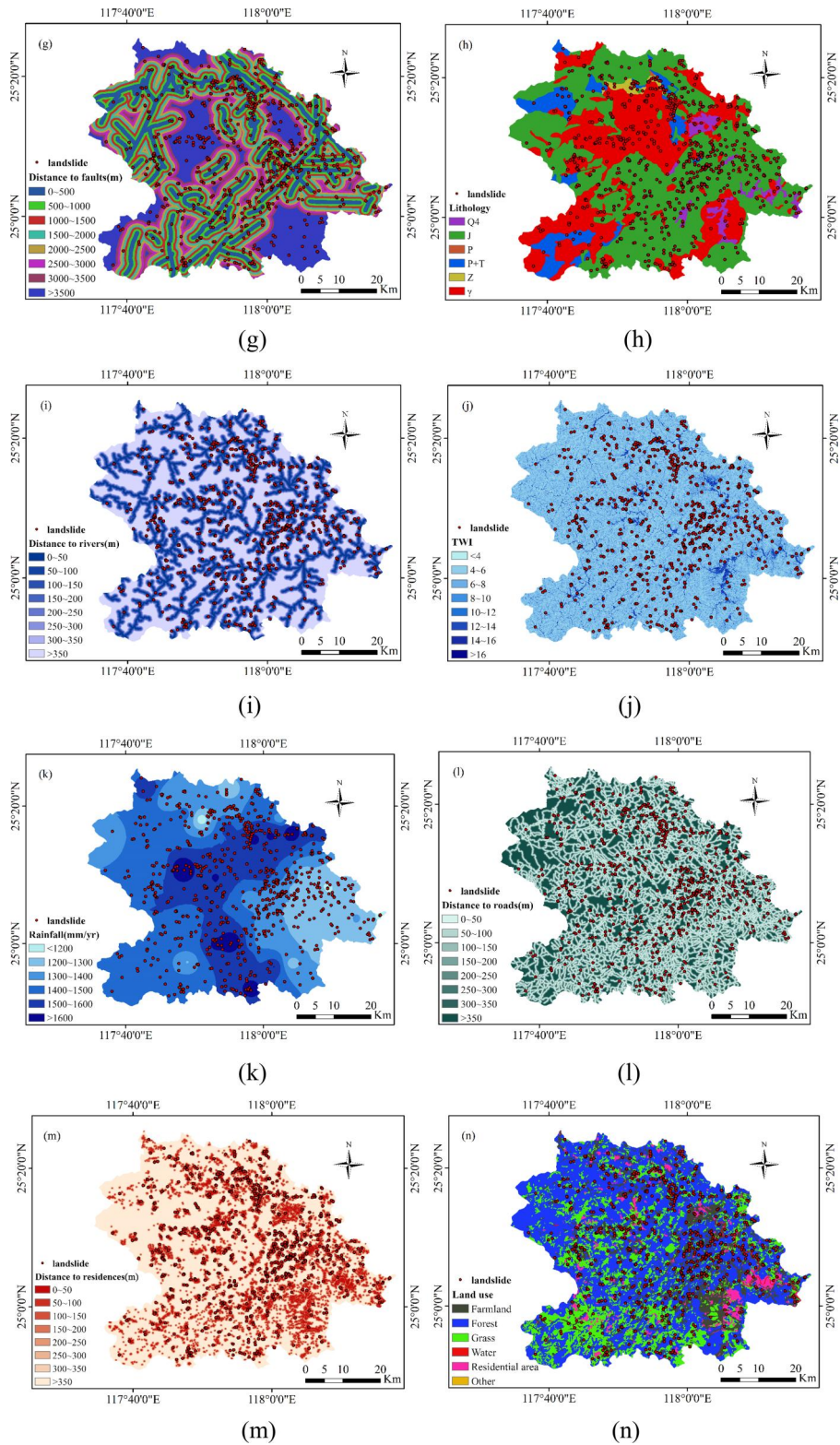


Figure 2. Continued.

map of Anxi County. Six different lithological units in the study area are alluvial double-layered soil bodies (Q4), volcanic rock formations (J), limestone formations (P), sandstone siltstone formations (P+), schistose mica-quartz schist formations (P + T), and granite formations (γ) (Figure 2(h)). As can be seen in Figure 2, the highest number of landslides occurred within the volcanic rock formation, followed by the granite formation.

One of the primary exogenous factors inducing landslides is water. The strength of the geotechnical body, particularly the weak surface, is reduced by groundwater infiltration and softening mud of water which enhances the risk of landslides on slopes. In this paper, the influence of surface water systems on the distribution of regional landslides is illustrated by the frequency of landslides occurring within a specific spatial distance from a surface water system. The distribution map of water systems was derived from the topographic map of the study area. The distance from the water system in the study area was divided into eight categories with a spacing of 50 m, considering the actual situation (Figure 2(i)).

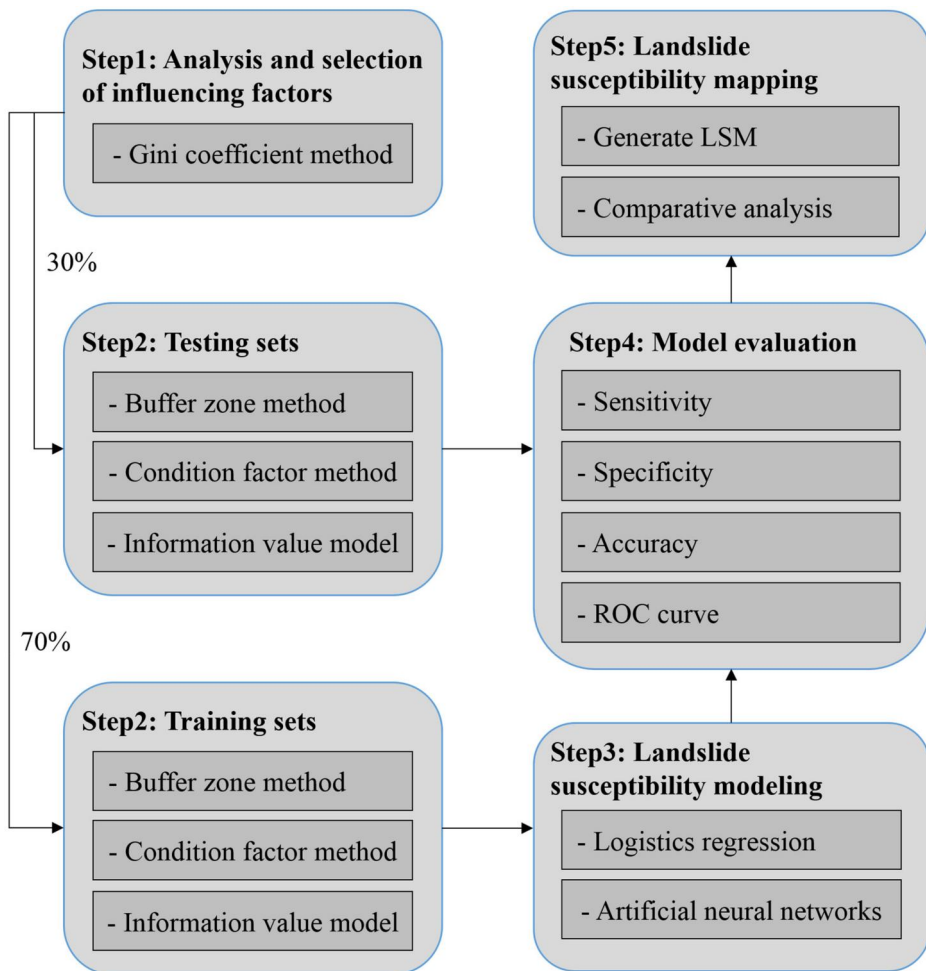
Terrain Wetness Index (TWI), which represents the spatial distribution of soil moisture, was derived from DEM and is classified into eight categories with a spacing of 2 m (Figure 2(j)).

Rainfall can weaken the rock and soil strength, and this weakening phenomenon in soil slopes is more evident after waterlogging. Similarly, the weakening in rock slopes causes a significant reduction in the shear strength which reduces the slope stability when the rock mass or the soft inclusions are more hydrophilic. This study used the five-year precipitation data of Anxi County from 2016 to 2020 to generate the average annual rainfall in recent years based on ArcGIS processing. The annual rainfall in Anxi County was classified into six categories with a spacing of 100 mm (Figure 2(k)).

Another significant factor impacting the formation of landslides is human activity. Due to the extensive slope cutting in road construction, which alters the slope structure and destroys the initial stress level of the geotechnical body, these progrades are highly susceptible to landslide hazards under specific condition. The geological hazard survey report of Anxi County was used to calculate the distance from the road, which is classified into eight categories with an interval of 50 m (Figure 2(l)). As the study area is hilly mountainous, most houses are built on the slopes, increasing slope loads and modifying the slope stress distribution. The frequency of landslide occurrence within a specific spatial distance from houses is used to illustrate the influence of houses on the distribution of regional landslides while examining the impacts of houses on slopes. The map of distance from the houses was derived from the house map in the geological hazard survey report of Anxi County. Eight categories of distance from houses were used in this investigation, each with a spacing of 50 m (Figure 2(m)). Anxi County has experienced many landslides due to land usage, industrial and agricultural development, and rapid infrastructure establishments. Land use of the study area was downloaded from the National Basic Geographic Information Centre and was divided into six categories, which are arable land, forest land, grassland, water, construction land, and others (Figure 2(n)). Further, the landslide evaluation factor map was resampled into a raster format with a spatial resolution of 30×30 m.

Table 1. Data sources of the landslide condition factor.

Condition factor	Data source	Variable type	Method to get the value in the slope unit
Altitude	Digital Elevation Model (30m)	Continues	Average value
Slope angle		Continues	Average value
Aspect		Categorical	Most frequent value
Plan curvature		Continues	Average value
Profile curvature		Continues	Average value
TWI	Geological survey map (1:50,000 scale)	Continues	Average value
Lithology		Categorical	Most frequent value
Distance to fault		Continues	Minimum value
Rainfall	Fujian Meteorological Bureau	Continues	Average value
NDVI	Landsat 8 OLI_TIRS image (2021-02-06 date)	Continues	Average value
Distance to roads	Third National Land Census	Continues	Minimum value
Distance to river		Continues	Minimum value
Distance to residences		Continues	Maximum value
Land use		Categorical	Most frequent value

**Figure 3.** Flowchart of the study.

3. Methodologies

As shown in Figure 3, the study is divided into five processes which are (1) selection of suitable condition factors by building a landslide susceptibility evaluation index system based on the Gini index, (2) selection of the corresponding number of landslide and non-landslide samples using three different types of methods and their classification into training and test sets with a ratio of 70:30, (3) building corresponding landslide susceptibility models using different training sets, (4) validation of different non-landslide selection method models by multiple statistical validation parameters, and (5) generation of landslide susceptibility zoning maps and their comparative analysis.

3.1. Preparation of non-landslide datasets

The identification of labeled data must be prioritized by supervised types of machine learning methods to produce predictions, as the accuracy of this data will determine the prediction results. Several studies have shown that selecting non-landslide data significantly impacts the final evaluation results during landslide susceptibility assessment. The present study compares the machine learning susceptibility models based on the three selection strategies, which are: (1) random generation of non-landslide points equal to the landslide data outside the buffers by generating landslide buffers of different distances and cropping the buffers in the study area, (2) selection of areas with low landslide impact in the landslide susceptibility assessment factor layer by combining the selected condition factors with the landslide samples and the random generation of non-landslide data in these areas and (3) random generation of non-landslide points using the medium-low susceptibility intervals, selected based on the mapping results of the Anxi County information value model.

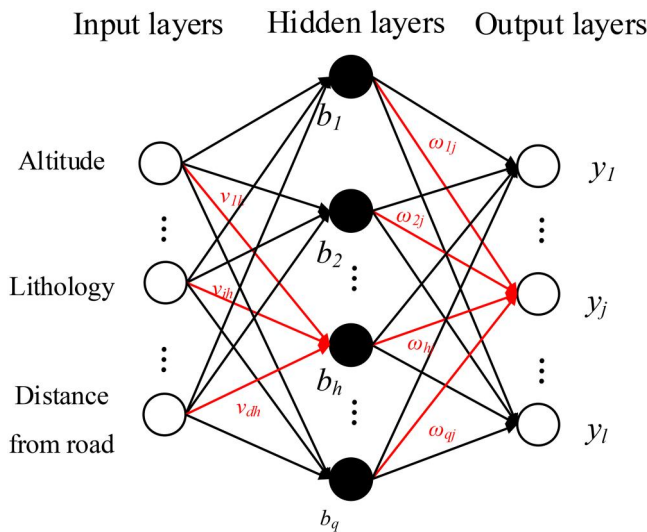


Figure 4. Artificial neural network structure.

The 821 landslide data used in the study area were randomly divided into two subsets with a ratio of 70:30. The exact number of non-landslide points was also generated randomly in the non-landslide area using the three methods mentioned above and were divided based on the 70:30 ratio. Finally, the landslide and nonlandslide points were assigned the values '1' and '0', respectively.

3.2. Information value model

A statistical technique known as the information value model based on probability theory has been widely used in landslide susceptibility mapping. These models use magnitudes of the entropy values during the occurrence of geological hazards to represent their probabilities. Information value defines the weight value of each classification unit. It is calculated as the ratio of the logarithm of the density of a landslides class to the density of landslides. The larger the information value, the more likely landslide hazards occur within the evaluation.

$$I_i = \log_2 \frac{N_i/N}{S_i/S} \quad (1)$$

where S is the total number of evaluation cells in the study area; N is the number of evaluation cells with landslides occurrence; S_i is the number of evaluation cells with an evaluation index ' i '; N_i is the number of cells with landslide occurrence and an evaluation index ' i '.

The following equation can calculate the total information value.

$$I = \sum_{i=1}^n I_i = \sum_{i=1}^n \log_2 \frac{N_i/N}{S_i/S} \quad (2)$$

3.3. Logistic regression model

A generalized linear regression technique known as the logistic regression model is often used for dichotomous variable analysis in statistics, mapping of results to within [0,1] by the activation function, and for the construction of the best-fit function between a dependent variable and multiple independent variables, which has been widely used in the spatial prediction of landslide hazards. The magnitude of landslide susceptibility is evaluated based on the probability value of landslide occurrence. The landslide susceptibility is defined as follows.

$$P = \frac{e^Y}{1 + e^Y} \quad (3)$$

where P and Y represent the probability of landslide occurrence and the linear regression equation, respectively. The latter can be defined using the following equation.

$$Y = \ln \left(\frac{P}{1-P} \right) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \alpha \quad (4)$$

where α , n , β_i , and x_i represent the intercept, the number of variables, the regression coefficient of variable x_i , and the variable factors controlling the landslide occurrence.

3.4. Artificial neural network

Artificial neural networks are frequently used in landslide sensitivity evaluation as they can mathematically simulate the biologic function of humans due to their fast convergence, extreme non-linearity, and high fault tolerance (Figure 4).

For a given training dataset (x, y) , the output expression of the neuron is

$$y = f\left(\sum_{i=1}^n \omega_i x_i + \alpha\right) \quad (5)$$

where f , w_i , x_i , and a indicates activation function, connection weight, input value, and deviation, respectively.

3.5. Selection of landslide condition factor by Gini index

Gini index, also called mean decrease impurity, is widely used by researchers to measure the importance of different condition factors. Gini index can be mathematically expressed as:

$$Gini(n, v_i) = \sum_{i=1}^m \frac{a_i}{f_s} I(d_{ui}) \quad (6)$$

where m donates the number of landslides at node n , f_s donates the number of input feature vectors used for training, the Gini impurity $I(d_{ui})$ represents the distribution

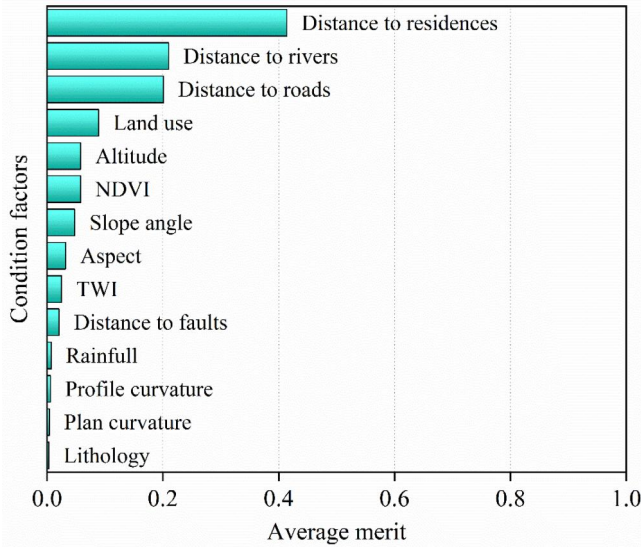


Figure 5. Average IG of conditioning factors.

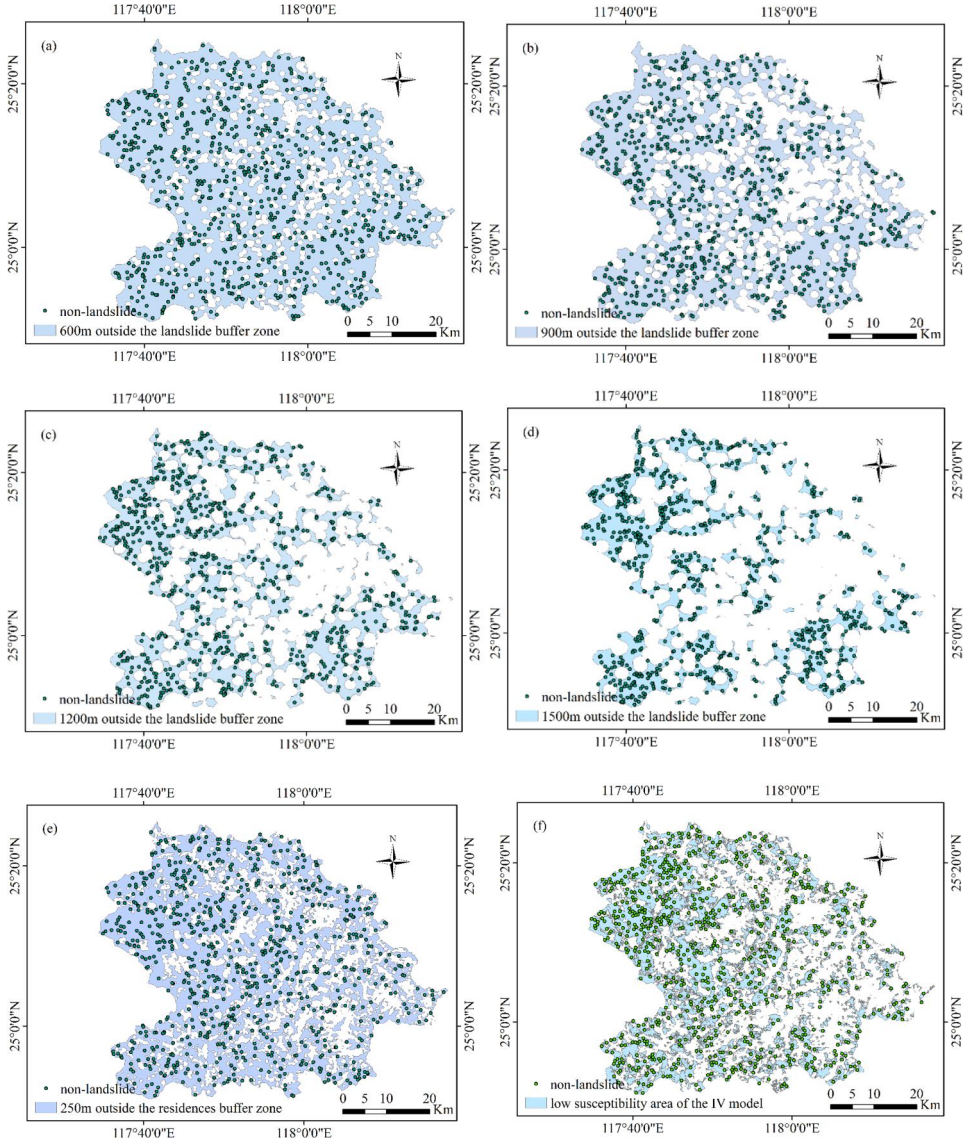


Figure 6. Non-landslide inventory map: (a), (b), (c), and (d) represent 600, 900, 1200, and 1500 m, respectively, outside the landslide buffer zone. (e) Non-landslide inventory map: 250 m outside the residences buffer zone. (f) Non-landslide inventory map: with very low and low susceptibility areas of the IV model.

of class labels in the node. For a feature variable $v_i \in V$ with m values at node n , $v_i = [u_1, u_2, \dots, u_m]$, the value of $I(d_{ui})$ can be computed as:

$$I(d_{ui}) = 1 - \sum_{i=0}^c \left(\frac{p_{c_i}}{a_i} \right)^2 \quad (7)$$

where p_{c_i} donates the number of samples with a value u_i belong to class c_i , a_i indicates the number of samples with a value u_i at that node n .

3.6. Accuracy evaluation and comparison

In this study, the sensitivity, specificity, accuracy and receiver operating characteristic curve (ROC) of landslide susceptibility model are assessed using each metric in the confusion matrix.

The sensitivity, which is the ratio of the number of samples predicted by the model to be in a positive class to the number of actual positive samples, is calculated using the following relation

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (8)$$

Specificity, which is the ratio of the number of samples predicted by the model to be true negative values to the total number of negative samples, is calculated using the following relation

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (9)$$

The accuracy of the model, the ratio of the number of accurate model predictions to the total number of predictions of the sample, is calculated using the following relation.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

where TP and TN accurately predict the results of landslides and non-landslides, respectively, while FP and FN incorrectly predict the results of landslides and non-landslides.

The receiver operating characteristic curve is the most commonly used quantitative evaluation technique for the prediction accuracy of landslide susceptibility model. The calculations and results of this method are simple and clear. Furthermore, the area under the curve (AUC) can be used to quickly test the accuracy of the predicted model. The value range of AUC is $0.5 \sim 1$, the closer the AUC value of a model is to 1, the higher the accuracy of the model.

4. Results and discussion

4.1. Landslide condition factor selection

In this study, the weights (AM) of 20 commonly used landslide impact factors are calculated using the Gini index, excluding the factors with $AM < 0$ and ranking the 14 factors with $AM > 0$ from largest to smallest. All the 14 selected factors have positive values ($AM > 0$) on landslide occurrence, as shown in [Figure 5](#). The distance to residences has the greatest value ($AM = 0.414$) due to the occurrence of most historical landslide events procured from Anxi County Natural Resources Bureau either in front of or behind the residential structures, followed by distance to rivers ($AM = 0.210$), distance to roads ($AM = 0.201$), land use ($AM = 0.089$), altitude ($AM =$

Table 2. Information values of landslide evaluation factors.

Conditioning factors	Classes	Percentage of landslide (%)	Percentage of domain (%)	I_i
Slope angle (°)	0–10	19.488	19.316	0.013
	10–20	47.503	37.337	0.347
	20–30	25.944	29.879	−0.204
	30–40	6.090	11.365	−0.900
	40–50	0.974	1.983	−1.025
	50–60	0	0.119	0
Altitude (m)	8–200	16.200	10.733	0.594
	200–400	23.630	14.976	0.658
	400–600	23.264	20.819	0.160
	600–800	23.630	26.961	−0.190
	800–1000	11.693	17.665	−0.595
	1000–1200	1.583	5.956	−1.911
	1200–1400	0.000	1.516	0.000
Distance to residences (m)	0–50	47.747	7.712	2.630
	50–100	28.989	8.693	1.738
	100–150	12.058	7.021	0.780
	150–200	4.385	5.909	−0.430
	200–250	3.167	6.675	−1.076
	250–300	0.853	5.928	−2.797
	300–350	0.731	5.477	−2.906
	350<	2.071	52.587	−4.667
TWI	< 4	1.462	3.398	−1.217
	4–6	57.856	58.979	−0.028
	6–8	22.655	22.470	0.012
	8–10	7.430	6.784	0.131
	10–12	3.532	3.546	−0.006
	12–14	5.481	3.138	0.805
	14–16	0.853	0.954	−0.162
	16<	0.731	0.732	−0.003
Plan curvature	−3.56~−0.64	0.244	0.707	−1.537
	−0.64~−0.37	1.583	3.980	−1.330
	−0.37~−0.19	12.667	15.941	−0.332
	−0.19~−0.02	25.457	23.339	0.125
	−0.02 ~ 0.14	36.419	29.685	0.295
	0.14 ~ 0.30	16.322	15.523	0.072
	0.30 ~ 0.54	6.090	8.502	−0.481
	0.54 ~ 3.32	1.218	2.027	−0.735
Profile curvature	−5.76~−0.91	0	0.255	0
	−0.91~−0.47	0.853	2.633	−1.627
	−0.47~−0.21	8.892	12.850	−0.531
	−0.21~−0.03	19.488	22.275	−0.193
	−0.03 ~ 0.14	39.708	32.655	0.282
	0.14 ~ 0.32	21.803	17.389	0.326
	0.32 ~ 0.62	9.013	7.968	0.178
	0.62 ~ 5.39	0.244	1.034	−2.085
Lithology	Q4	3.532	3.240	0.125
	J	55.664	54.240	0.037
	P	0	0.033	0
	P + T	5.359	7.237	−0.433
	Z	3.532	0.908	1.960
	γ	31.912	34.343	−0.106
Distance to rivers (m)	0–50	25.822	11.813	1.128
	50–100	20.341	9.786	1.056
	100–150	16.565	10.141	0.708
	150–200	11.449	8.332	0.458
	200–250	8.526	8.534	−0.001
	250–300	4.994	7.812	−0.646
	300–350	4.141	6.924	−0.742
	350	8.161	36.657	−2.167
Rainfall (mm/yr)	<1200	0.731	0.391	0.904
	1200–1300	13.398	13.773	−0.040

(continued)

Table 2. Continued.

Conditioning factors	Classes	Percentage of landslide (%)	Percentage of domain (%)	I_i
Distance to roads (m)	1300–1400	23.752	20.008	0.247
	1400–1500	29.233	38.186	−0.385
	1500–1600	27.893	24.485	0.188
	>1600	4.994	3.158	0.661
	0–50	35.323	21.780	0.698
	50–100	27.040	17.823	0.601
	100–150	18.514	15.249	0.280
	150–200	6.456	9.054	−0.488
	200–250	5.603	9.444	−0.753
	250–300	4.750	7.005	−0.560
Land use	300–350	0.974	4.422	−2.182
	350–2429	1.340	15.223	−3.506
	Farmland	36.297	16.391	1.147
	Forest	33.374	54.142	−0.698
	Grass	24.361	24.592	−0.014
	Water	0.122	0.374	−1.617
	Residential area	5.847	4.495	0.379
NDVI	Other	0	0.006	0
	−1–0.09	3.776	2.437	0.632
	0.09–0.17	24.726	7.530	1.715
	0.17–0.24	30.085	16.500	0.867
	0.24–0.29	21.924	25.896	−0.240
	0.29–0.36	13.886	29.704	−1.097
	0.36–0.57	5.603	17.934	−1.678
Distance to faults (m)	0–500	20.828	17.099	0.285
	500–1000	20.463	16.123	0.344
	1000–1500	14.251	14.145	0.011
	1500–2000	13.886	11.907	0.222
	2000–2500	8.892	9.850	−0.148
	2500–3000	6.212	7.418	−0.256
	3000–3500	3.898	5.420	−0.476
	>3500	11.571	18.038	−0.641
Aspect	Flat	0	0.009	0
	North	9.866	12.640	−0.357
	Northeast	14.251	13.536	0.074
	East	14.738	12.053	0.290
	Southeast	14.495	13.042	0.152
	South	14.616	12.575	0.217
	Southwest	13.886	12.857	0.111
	West	8.526	11.207	−0.394
	Northwest	9.622	12.081	−0.328

0.058), NDVI (AM = 0.058), slope angle (AM = 0.048), aspect (AM = 0.032), and TWI (AM = 0.025), distance to faults (AM = 0.021), rainfall (AM = 0.007), profile curvature (AM = 0.006), plan curvature (AM = 0.004) and lithology (AM = 0.003).

4.2. Non-landslide data generation

4.2.1. Landslide buffer generates non-landslide data

Based on previous landslide data, landslide buffers were created as 600, 900, 1200, and 1500 m. The GIS cropping tool was used to crop several landslide buffer distances in the base map, which is the regional map of Anxi County, while non-landslide data were randomly generated from the cropped area as the base map (Figure 6(a–d)).

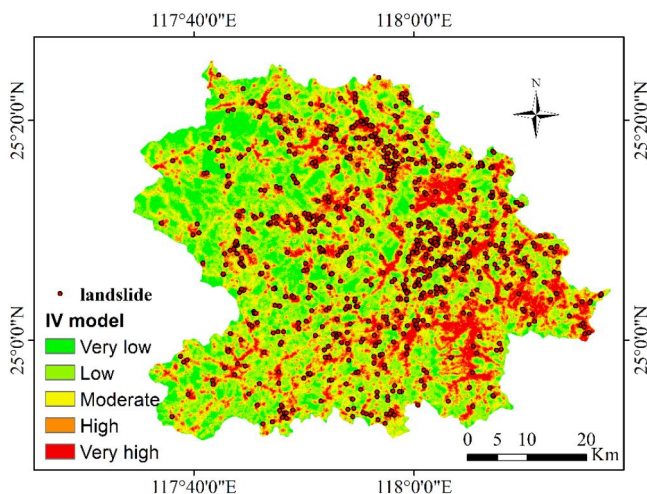


Figure 7. Landslide susceptibility map generated by IV model.

Table 3. Condition factor logistic regression coefficients.

Variables	Model A ₁	Model A ₂	Model A ₃	Model A ₄	Model A ₅	Model A ₆
Slope angle	−0.234	−0.219	−0.158	−0.111	0.120	1.042
Altitude	0.113	0.010	0.041	−0.022	0.273	1.524
Distance to residences	2.239	2.243	2.408	0.810	5.184	5.623
TWI	0.338	0.271	0.385	0.474	0.613	0.699
Plan curvature	−0.080	0.056	−0.150	−0.023	−0.196	−0.324
Profile curvature	−0.229	0.059	−0.019	−0.054	−0.188	−0.330
Lithology	0.281	0.076	0.280	0.183	0.338	0.319
Distance to rivers	0.491	0.547	0.509	0.284	−0.072	1.310
Rainfall	−0.323	−0.095	−0.517	−0.552	−0.468	−0.890
Distance to roads	0.143	0.272	0.251	0.118	0.653	1.385
Land use	−0.015	−0.335	−0.271	−0.047	−0.208	0.200
NDVI	0.198	0.235	0.411	0.228	0.114	0.993
Distance to faults	0.372	0.475	0.679	0.335	0.307	0.729
Aspect	0.121	0.205	0.252	0.132	0.411	0.656
Bias	0.285	0.278	0.270	−7.127	−0.805	−2.256

4.2.2. Conditioning factor generation for non-landslide data

In order to avoid incorrect choosing of locations with significant historical landslide occurrence while generating non-landslide data, the indicator factor layer grading area used to generate non-landslide data must ensure fewer historical landslides in the area. To this end, the following non-landslide area principles are developed: (a) The regions selected as non-landslide areas should have a relatively low proportion of historical statistical landslide data. (b) The extent of non-landslide areas needs to be significant to prevent the clustering of non-landslide data and to generate a sufficient comprehensive representation of other areas. Based on the principles mentioned above and the statistical analysis of the evaluation factors, the areas satisfying the first condition were identified, which include areas with (i) slope above 40° , (ii) elevation greater than 1000 m, (iii) distance from houses above 250 m, (iv) TWI above 14, (v) limestone formation as stratigraphic lithology, (vi) rainfall below 1200 mm, (vii) distance from roads above 300 m, (viii) land use type of water and others. Out of these areas satisfying the first condition, the areas meeting the second condition were

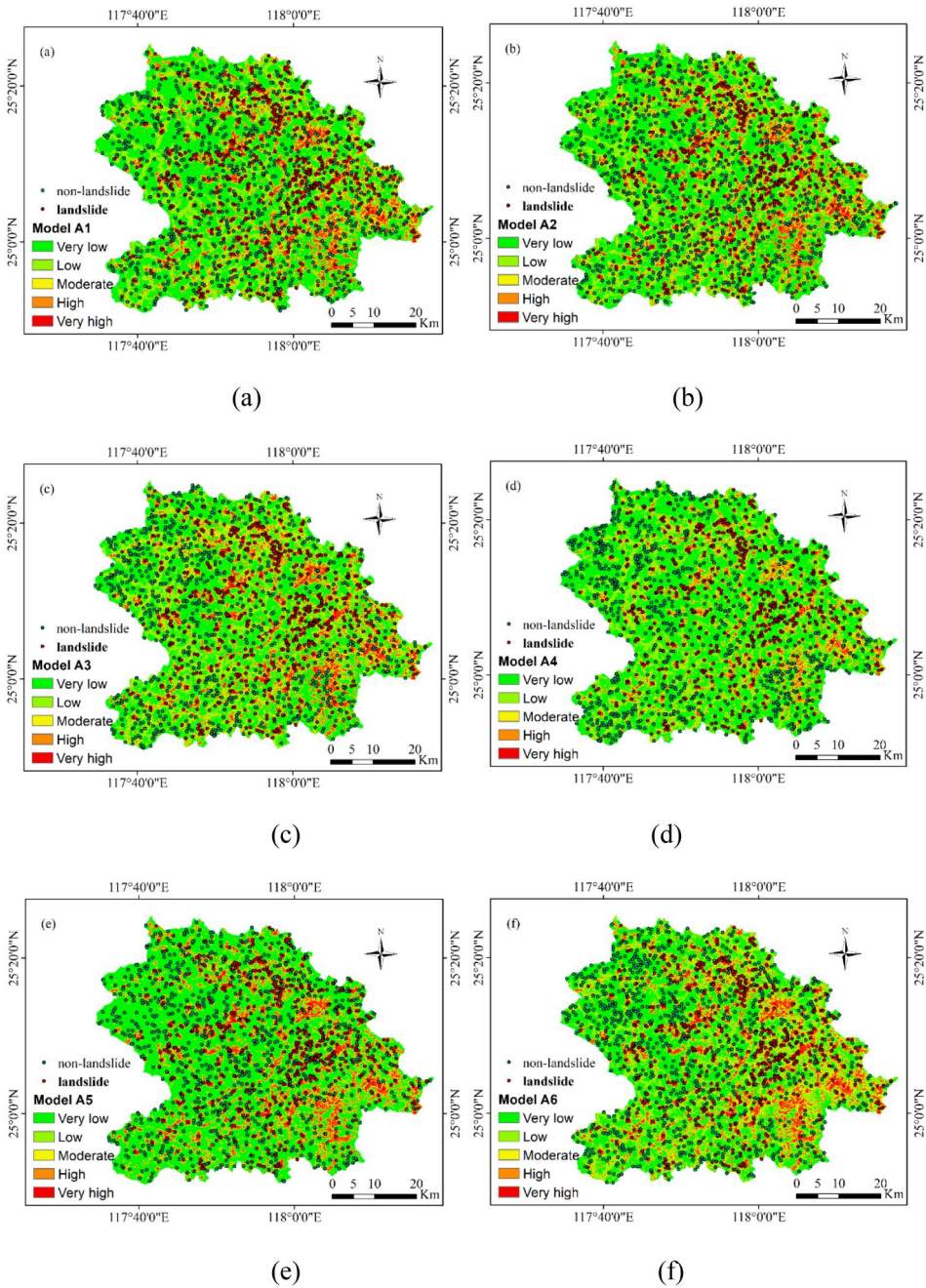


Figure 8. Landslide susceptibility maps using: (a) model A1, (b) model A2, (c) model A3, (d) model A4, (e) model A5, and (f) model A6.

identified, which include a category with (i) 64% of the area located more than 250 m away from houses and (ii) 20% of the area located more than 300 m away from roads. The percentage of landslides in the former category is only 3.5%, which is half of the

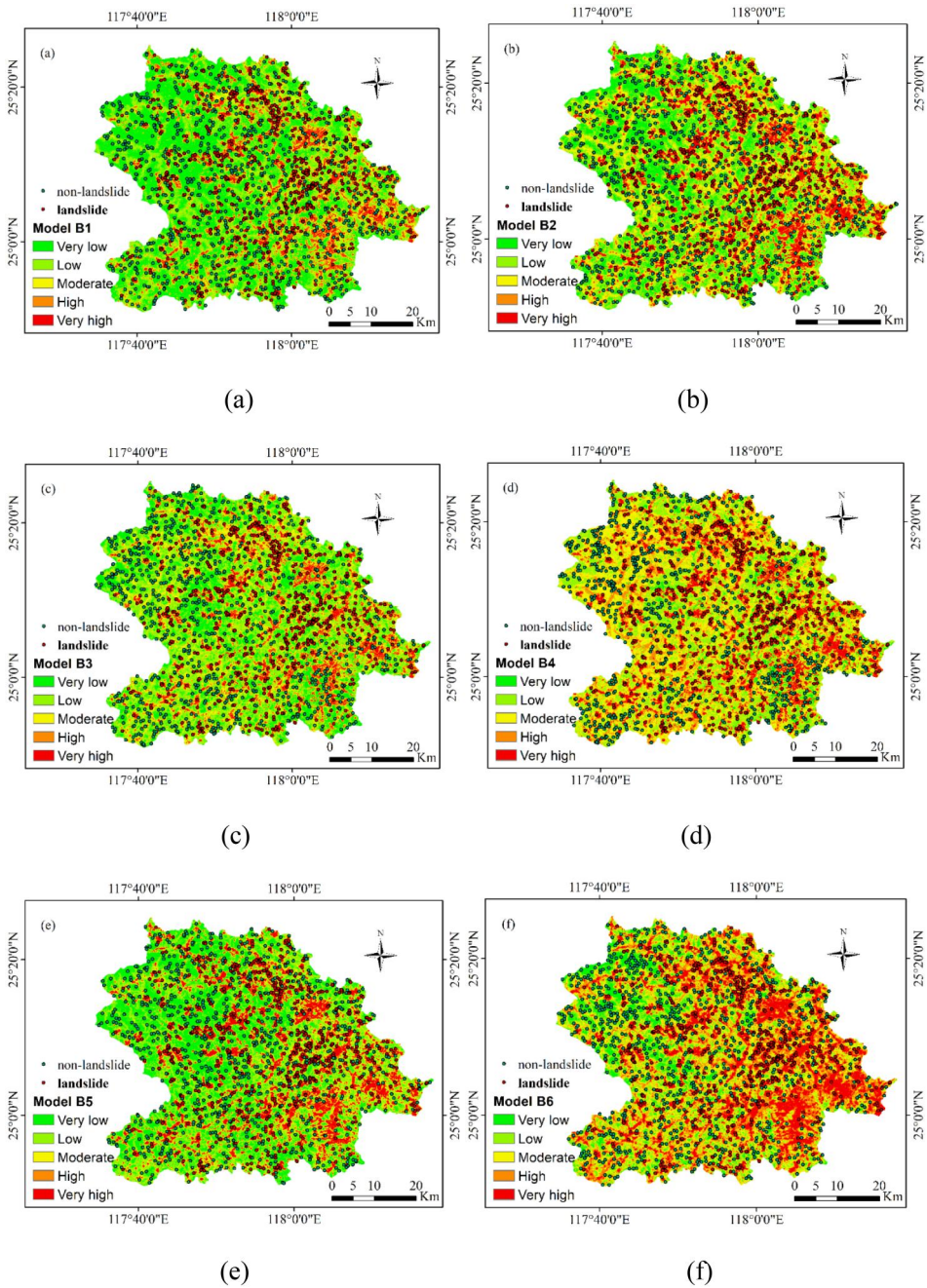


Figure 9. Landslide susceptibility maps using: (a) model B1, (b) model B2, (c) model B3, (d) model B4, (e) model B5, and (f) model B6.

landslide occurrence in the latter category. In contrast, the former area is three times larger than the latter. In conclusion, 821 randomly generated non-landslide points were created in the area over 250 meters from the house (Figure 6(e)).

4.2.3. Information value model generation for non-landslide data

The non-landslide data generated based on the area selected by a single layer may still appear to be overly dominated by a single factor, even though there are two principles for selecting non-landslide areas. However, such situations can be overcome by considering the results of the partitioning of the Anxi County information value model for selecting lower and low susceptibility intervals as non-landslide areas and the subsequent random generation of the non-landslide points. The selected landslide evaluation factor layers were analyzed by extracting the numbers of landslides and zoning area for each factor grading interval using GIS. Further, the ratio of the number of landslides to the number of rasters for the index factor interval was calculated to derive the information value for each interval based on the information value model formula. The calculation results are shown in Table 2.

The superposition analysis was performed based on the information value calculation formula using the weighted superposition tool in GIS to generate the landslide susceptibility map of Anxi County, which was further divided into five susceptibility classes using the natural interruption point method (Figure 7). Finally, an equal number of non-landslide points were randomly generated using both lower susceptibility zone and the low susceptibility zone of the information value model as non-landslide areas (Figure 6(f)).

4.3. Different non-landslide data based on logistic regression models

The dataset required to build the model was generated by combining the landslide sample set with the data extraction of the 14 condition factors based on the non-landslide sample set generated by the three different selection methods mentioned above. Further, the dataset was divided into training and test sets in a 70:30 ratio using a stratified sampling method. Finally, multiple logistic regression models were developed, and the condition factor regression coefficients are listed in Table 3, where Model A1, A2, A3, and A4, are logistic regression models of various buffer zones of landslides such as 600, 900, 1200, and 1500 m, respectively, while A5 and A6 are logistic regression models for more than 250 m away from the house and that for low information prone area.

Further, the susceptibility index of Anxi County was determined based on the logistic regression model outside the buffer distance of different landslide hazards by combining the coefficients of the logistic regression model in Equation (4) using the raster calculator. Based on the natural interruption method, the susceptibility index was classified into five categories: very low susceptibility zone, low susceptibility zone, medium susceptibility zone, high susceptibility zone, and very high susceptibility zone, as shown in Figure 8(a–f).

4.4. Different non-landslide data based on artificial neural network

The artificial neural network model used in the study had an input layer with 14 evaluation factors, nine hidden layers, and two output layers. The artificial neural

network model uses heuristic algorithm to self-optimize parameters in the training process, and the momentums were set as 0.5, 0.9, and 0.99, and the learning rates were set as 0.01 and 0.1. The artificial neural network model parameters for training on the other datasets in this study include a learning rate of 0.01, and a model momentum of 0.5, with 300 cycles. The connection weights were calculated between each neuron in the input and hidden layers and between the hidden layer and the output layer. The landslide susceptibility indices were calculated by combining the neural network weights with the activation function. They were classified into five categories using the reclassification tool (Figure 9). Model B1, B2, model B3, and model B4 are ANN models for landslide buffer zones 600 m, 900 m, 1200, and 1500, respectively, while model B5 and model B6 are ANN models for a distance of 250 m or more from a house, and that for a low information-prone zone, respectively.

4.5. Model validation

4.5.1. Validation of logistic regression model

The number of zonal rasters for each model is different due to the use of the natural interruption point method for the classification of landslide susceptibility indices, which results in different areas within the zone, affecting the historical landslide data from the zonal statistics. Several statistical parameters were introduced to compare the advantages and disadvantages of the non-landslide samples selected by each method and evaluate each model's accuracy. The estimated evaluation indices are shown in Table 4.

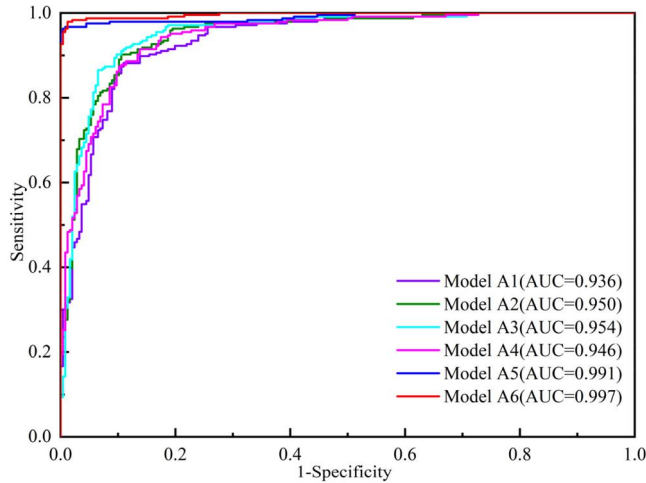
As evident from Table 4, the best performance in landslide unit classification is shown by model A6 (sensitivity = 98.0%), followed by model A5 (sensitivity = 96.7%), models A2 and A3 (sensitivity = 90.2%), model A1 (sensitivity = 89.8%), and model A4 (sensitivity = 89.4%). Similarly, the best performance in the classification of non-landslide units was shown by model A5 (specificity = 99.2%), followed by model A6 (specificity = 98.8%), model A3 (specificity = 89.4%), model A2 (specificity = 89.0%), model A4 (specificity = 86.6%), and model A1 (specificity = 85.8%). The overall best performance in the classification of landslide and non-landslide units is shown by model A6 with the highest accuracy of 98.4%, followed by model A5 (accuracy = 98.0%), model A3 (accuracy = 89.8%), model A2 (accuracy = 89.6%), model A4 (accuracy = 88.0%), and model A1 (accuracy = 87.8%).

The ROC curves corresponding to the six models were generated to more intuitively compare the prediction accuracy and stability among the models. The ROC curves are shown in Figure 10, and the area under the curve (AUC) is given in Table 5. Based on the LR model, all the non-landslide sampling strategies show high performance ($AUC > 0.9$). However, the best performance is shown by model A6 ($AUC = 0.997$), followed by model A5 ($AUC = 0.991$), model A3 ($AUC = 0.954$), model A2 ($AUC = 0.950$), model A4 ($AUC = 0.946$) and model A1 ($AUC = 0.936$).

Overall, all the non-landslide sampling methods are acceptable for mapping landslide susceptibility in the study area. Based on the study's findings, it can be concluded that the non-slippery samples generated using the low susceptibility area of the information value model in the LR model exhibited the best performance.

Table 4. Accuracy validation parameters of logistic regression models.

Parameters	Model A1	Model A2	Model A3	Model A4	Model A5	Model A6
Sensitivity	0.898	0.902	0.902	0.894	0.967	0.980
Specificity	0.858	0.890	0.894	0.866	0.992	0.988
Accuracy	0.878	0.896	0.898	0.880	0.980	0.984

**Figure 10.** ROC curves of the six logistic regression models.

4.5.2. Validation of artificial neural network model

Statistical parameters are used to validate the accuracy of the six model sets obtained and to compare the advantages and disadvantages of the three non-landslide sampling methods under the ANN-based model. The validation results are shown in Table 6.

As evident from Table 6, the best performance in landslide unit classification was shown by model B6 (sensitivity = 98.0%), followed by model B5 (sensitivity = 95.9%), model B4 (sensitivity = 91.6%), and models B1, B2 and B3 (sensitivity = 90.7%). Similarly, the best performance in the classification of non-landslide units was shown by model B5 (specificity = 98.8%), followed by model B6 (specificity = 98.4%), model B3 (specificity = 88.6%), model B2 (specificity = 86.2%), model B4 (specificity = 85.8%) and model B1 (specificity = 84.6%). However, the overall best performance in the classification of landslide and non-landslide units was shown by model B6, which had the highest accuracy of 98.2%, followed by model B5 (accuracy = 97.0%), model B3 (accuracy = 89.6%), model B4 (accuracy = 88.7%), model B2 (accuracy = 88.4%), and model B1 (accuracy = 87.6%).

The general performance of the landslide models using ROC curves and validation data is shown in Table 7 and Figure 11. Based on ANN models, all the non-slippery slope sampling strategies are observed to exhibit high performance ($\text{ANN} > 0.9$). The highest performance was shown by Model B6 ($\text{AUC} = 0.995$), followed by model B5 ($\text{AUC} = 0.989$), model B3 ($\text{AUC} = 0.956$), model B2 ($\text{AUC} = 0.952$), model B4 ($\text{AUC} = 0.946$) and model B1 ($\text{AUC} = 0.942$).

Overall, all the non-landslide sampling methods are suitable for mapping landslide susceptibility in the study area. Based on the above analysis, it can be concluded that

Table 5. AUC analysis for the six logistic regression models.

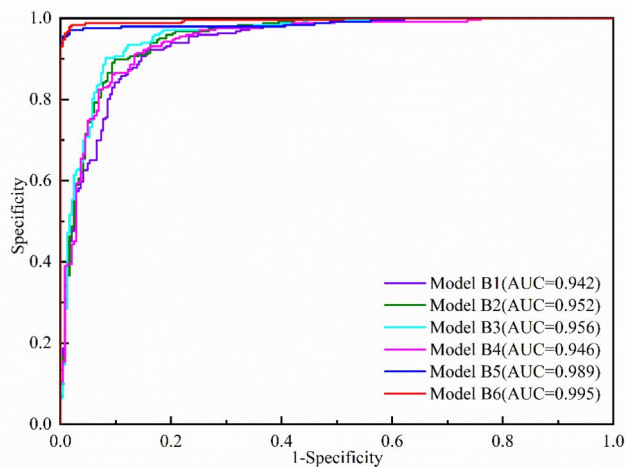
Models	AUC	SE	95% CI
Model A1	0.936	0.011	0.915–0.958
Model A2	0.950	0.010	0.931–0.968
Model A3	0.954	0.009	0.935–0.972
Model A4	0.946	0.010	0.926–0.965
Model A5	0.991	0.004	0.983–0.998
Model A6	0.997	0.002	0.993–1.000

Table 6. Accuracy validation parameters of artificial neural network models.

Parameters	Model B1	Model B2	Model B3	Model B4	Model B5	Model B6
Sensitivity	0.907	0.907	0.907	0.916	0.959	0.980
Specificity	0.846	0.862	0.886	0.858	0.988	0.984
Accuracy	0.876	0.884	0.896	0.887	0.974	0.982

Table 7. AUC analysis for the six artificial neural network models.

Models	AUC	SE	95%CI
Model B1	0.942	0.010	0.922–0.961
Model B2	0.952	0.009	0.935–0.970
Model B3	0.956	0.009	0.938–0.973
Model B4	0.946	0.010	0.926–0.965
Model B5	0.989	0.005	0.980–0.998
Model B6	0.995	0.002	0.991–1.000

**Figure 11.** ROC curves of the six artificial neural network models.

the use of the low susceptibility area of the information value model showed the best performance in the non-landslide sampling method in the ANN model.

4.5.3. Compared with models from other studies

For the selection strategy of non-landslide samples, some researchers used different methods to conduct similar research (Table 8). For example, Wang et al.(2022) used a special non-landslide sample selection method and a logistic regression algorithm to

Table 8. Comparison of model prediction accuracy in different non-landslide point selection strategies (ROC)

Source	Condition factors	Model	non-landslide sample selection strategies	ROC
Wang et al.(2022)	Slope, Curvature, Terrain relief, NDVI, Drainage density, Curve number, Soil type, Lithology, GDP, Population density, Average annual precipitation over 10mm/d, Landslide density	Logistic regression	Selection in the entire area	Training 0.845 Validating 0.830
			Selection in the mountain area	Training 0.736 Validating 0.732
Li et al.(2021)	Geotechnical, Distance to Fault, Slope, Aspect, NDVI, Land-use, Distance to Road, Distance to River, Rainfall	Support vector machines	Random sampling method	Training 0.845 Validating 0.830
			Buffer sampling method	Training 0.845 Validating 0.830
Chang et al.(2023)	Altitude difference, Aspect, Relief amplitude, Distance to river, Normalized difference built-up index (NDBI)	Logistic regression	Information value method	Training 0.845 Validating 0.830
			Selected number N = 1	0.714
			Selected number N = 10	0.775
			Selected number N = 100	0.780
			Selected number N = 500	0.783
			Selected number N = 1000	0.815
			Selected number N = 5000	0.787
		Support vector machines	Selected number N =1	0.711
			Selected number N = 10	0.768
			Selected number N = 100	0.770
			Selected number N = 500	0.773
			Selected number N = 1000	0.775
			Selected number N = 5000	0.775
This study	Slope angle, Altitude, Aspect, NDVI, Plan curvature, Profile curvature, Distance to faults, Lithology, Distance to rivers, TWI, Rainfall, Distance to roads, Distance to residences, Land use	Logistic regression	Landslide buffers 600 m	0.936
			Landslide buffers 900 m	0.950
			Landslide buffers 1200 m	0.954
			Landslide buffers 1500 m	0.946
			250 m outside the residences buffer	0.991
			Information value method	0.997
		Artificial neural networks	Landslide buffers 600 m	0.942
			Landslide buffers 900 m	0.952
			Landslide buffers 1200 m	0.956
			Landslide buffers 1500 m	0.946
			250 m outside the residences buffer	0.989
			Information value method	0.995

model landslide susceptibility, and finally found that the accuracy of the model sampled only in the mountainous area ($AUC = 0.732$) was lower than that of the model sampled in the whole research area ($AUC = 0.830$). Li et al. (2021) extracted non-landslide samples using random, buffer, and information value methods, respectively, and applied the obtained non-landslides to the SVM model in Landslide Hazard Mapping in the western area of Tumen City. The results show that the prediction accuracy of the three methods is above 80%. Chang et al. (2023) constructed a logistic regression and support vector machine model based on slope units using 16 condition factors in Chongyi County, and discussed the effect of the number of sample extractions on non-landslide uncertainty. The results show that the accuracy and rationality of the evaluation results of landslide susceptibility increase with the increase of the number of non-landslide samples.

On the basis of the above research, this paper improves some previous methods and puts forward a non-landslide sample selection method based on the buffer method, condition factor method and information value method. Then, the landslide susceptibility model of the study area are established by logistic regression and artificial neural network respectively. The calculation results show that the accuracy of the proposed method is obviously higher than that of the previous methods. It must be recognized that this could be related to the differences between the study area and its historical landslide samples.

5. Conclusion

This paper presents a novel approach to developing three non-landslide sampling methods, namely, different landslide distance buffers, condition factors, and information value models, to examine the influence of non-landslide point selection strategies on the results of landslide susceptibility evaluation. Further, the performance of three non-landslide sampling methods in landslide susceptibility modeling and mapping was evaluated using logistic regression model and artificial neural network model. The main conclusions are as follows:

1. Non-landslide sample selection significantly impacts the prediction performance of landslide susceptibility, with enhanced accuracy of landslide susceptibility mapping while adopting a reasonable sample selection strategy for non-landslide points.
2. The AUC values of all the models developed in this paper are greater than 90%, indicating the high accuracy and reliability of the machine learning susceptibility evaluation models constructed by combining non-landslides sample selection methods.
3. The approach employing a low susceptibility area of the information value model as the non-landslide generation area has the highest accuracy compared to other three major existing methods for non-landslide selection. It minimizes the probability of the selected non-landslide occurrences in high-risk areas, enhancing landslide prediction accuracy.

4. Although the current study optimized the non-landslide selection strategy with the highest prediction accuracy through comparison, it lacks the exploration of the influence of parameters on the evaluation results due to conditions such as insufficient spatial and temporal data. In view of this, it is suggested that the dynamic risk assessment of regional landslides can be carried out based on time series data and vulnerability assessment in the future, and the influence of parameters in non-landslide selection strategy on the accuracy of the evaluation model can be further discussed.

Disclosure statement

All authors disclosed no relevant relationships, and authors have no conflict of interest to declare.

Funding

The authors would like to acknowledge the financial support from the National Natural Science Foundation of China (No. U2005205, No. 42007235) and the Science and Natural Science Foundation of Fujian Province (No. 2023J01423).

Data availability

Data will be made available on request.

References

- Akinci H, Yavuz Ozalp A. **2021**. Landslide susceptibility mapping and hazard assessment in Artvin (Turkey) using frequency ratio and modified information value model. *Acta Geophys.* 69(3):725–745. doi: [10.1007/s11600-021-00577-7](https://doi.org/10.1007/s11600-021-00577-7).
- Akinci H. **2022**. Assessment of rainfall-induced landslide susceptibility in Artvin, Turkey using machine learning techniques. *J Afr Earth Sci.* 191:104535. doi: [10.1016/j.jafrearsci.2022.104535](https://doi.org/10.1016/j.jafrearsci.2022.104535).
- Azarafza M, Azarafza M, Akgün H, Atkinson PM, Derakhshani R. **2021**. Deep learning-based landslide susceptibility mapping. *Sci Rep.* 11(1):24112. doi: [10.1038/s41598-021-03585-1](https://doi.org/10.1038/s41598-021-03585-1).
- Chen W, Pourghasemi HR, Panahi M, Kornejady A, Wang J, Xie X, Cao S. **2017**. Spatial prediction of landslide susceptibility using an adaptive neuro-fuzzy inference system combined with frequency ratio, generalized additive model, and support vector machine techniques. *Geomorphology.* 297:69–85. doi: [10.1016/j.geomorph.2017.09.007](https://doi.org/10.1016/j.geomorph.2017.09.007).
- Chen W, Xie X, Peng J, Shahabi H, Hong H, Bui DT, Duan Z, Li S, Zhu A-X. **2018**. GIS-based landslide susceptibility evaluation using a novel hybrid integration approach of bivariate statistical based random forest method. *Catena.* 164:135–149. doi: [10.1016/j.catena.2018.01.012](https://doi.org/10.1016/j.catena.2018.01.012).
- Chong XU. **2010**. GIS platform and certainty factor analysis method based Wenchuan earthquake-induced landslide susceptibility evaluation. *J Eng Geology.* 18(1):15.
- Constantin M, Bednarik M, Jurchescu MC, Vlaicu M. **2011**. Landslide susceptibility assessment using the bivariate statistical analysis and the index of entropy in the Sibiciu Basin (Romania). *Environ Earth Sci.* 63(2):397–406. doi: [10.1007/s12665-010-0724-y](https://doi.org/10.1007/s12665-010-0724-y).
- Chang Z, Catani F, Huang F, Liu G, Meena SR, Huang J, Zhou C. **2023**. Landslide susceptibility prediction using slope unit-based machine learning models considering the heterogeneity

- of conditioning factors. *J Rock Mech Geotech Eng.* 15(5):1127–1143. doi: [10.1016/j.jrmge.2022.07.009](https://doi.org/10.1016/j.jrmge.2022.07.009).
- Dou J, Yunus AP, Merghadi A, Shirzadi A, Nguyen H, Hussain Y, Avtar R, Chen Y, Pham BT, Yamagishi H, et al. 2020. Different sampling strategies for predicting landslide susceptibilities are deemed less consequential with deep learning. *Sci Total Environ.* 720:137320. doi: [10.1016/j.scitotenv.2020.137320](https://doi.org/10.1016/j.scitotenv.2020.137320).
- Gaidzik K, Ramírez-Herrera MT. 2021. The importance of input data on landslide susceptibility mapping. *Sci Rep.* 11(1):19334. doi: [10.1038/s41598-021-98830-y](https://doi.org/10.1038/s41598-021-98830-y).
- Hong H, Ilia I, Tsangaratos P, Chen W, Xu C. 2017. A hybrid fuzzy weight of evidence method in landslide susceptibility analysis on the Wuyuan area, China. *Geomorphology.* 290:1–16. doi: [10.1016/j.geomorph.2017.04.002](https://doi.org/10.1016/j.geomorph.2017.04.002).
- Hong H, Tsangaratos P, Ilia I, Liu J, Zhu A, Chen W. 2018. Application of fuzzy weight of evidence and data mining techniques in construction of flood susceptibility map of Poyang County, China. *Sci Total Environ.* 625:575–588. doi: [10.1016/j.scitotenv.2017.12.256](https://doi.org/10.1016/j.scitotenv.2017.12.256).
- Huang Y, Zhao L. 2018. Review on landslide susceptibility mapping using support vector machines. *Catena.* 165:520–529. doi: [10.1016/j.catena.2018.03.003](https://doi.org/10.1016/j.catena.2018.03.003).
- Ilia I, Tsangaratos P. 2016. Applying weight of evidence method and sensitivity analysis to produce a landslide susceptibility map. *Landslides.* 13(2):379–397. doi: [10.1007/s10346-015-0576-3](https://doi.org/10.1007/s10346-015-0576-3).
- Kavzoglu T, Sahin EK, Colkesen I. 2014. Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. *Landslides.* 11(3):425–439. doi: [10.1007/s10346-013-0391-7](https://doi.org/10.1007/s10346-013-0391-7).
- Kayastha P, Dhital MR, De Smedt F. 2013. Application of the analytical hierarchy process (AHP) for landslide susceptibility mapping: a case study from the Tinau watershed, west Nepal. *Comput Geosci.* 52:398–408. doi: [10.1016/j.cageo.2012.11.003](https://doi.org/10.1016/j.cageo.2012.11.003).
- Lee S, Choi J. 2004. Landslide susceptibility mapping using GIS and the weight-of-evidence model. *Int J Geograph Inform Sci.* 18(8):789–814. doi: [10.1080/13658810410001702003](https://doi.org/10.1080/13658810410001702003).
- Li X, Cheng J, Yu D, Han Y. 2021. Research on non-landslide selection method for landslide hazard mapping.
- Lima P, Steger S, Glade T. 2017. Comparison of non-landslide sampling strategies to counteract inventory-based biases within national-scale statistical landslide susceptibility models. *EGU General Assembly Conference Abstracts* 13523.
- Mohammady M, Pourghasemi HR, Pradhan B. 2012. Landslide susceptibility mapping at Golestan Province, Iran: a comparison between frequency ratio, Dempster-Shafer, and weights-of-evidence models. *J Asian Earth Sci.* 61:221–236. doi: [10.1016/j.jseae.2012.10.005](https://doi.org/10.1016/j.jseae.2012.10.005).
- Onagh M, Kumra VK, Rai PK. 2012. Landslide susceptibility mapping in a part of Uttarkashi district (India) by multiple linear regression method. *Int J Geology, Earth Environ Sci.* 2: 102–120.
- Peng L, Niu R, Huang B, Wu X, Zhao Y, Ye R. 2014. Landslide susceptibility mapping based on rough set theory and support vector machines: a case of the Three Gorges area, China. *Geomorphology.* 204:287–301. doi: [10.1016/j.geomorph.2013.08.013](https://doi.org/10.1016/j.geomorph.2013.08.013).
- Pham BT, Pradhan B, Bui DT, Prakash I, Dholakia MB. 2016. A comparative study of different machine learning methods for landslide susceptibility assessment: a case study of Uttarakhand area (India). *Environ Modell Software.* 84:240–250. doi: [10.1016/j.envsoft.2016.07.005](https://doi.org/10.1016/j.envsoft.2016.07.005).
- Pradhan, B, Xu, C, Dieu, T, et al. 2017. Rainfall-induced landslide susceptibility assessment at the Chongren area (China) using frequency ratio, certainty factor, and index of entropy. *Geocarto Int.* 32(2):139–154.
- Pradhan B, Lee S. 2010. Regional landslide susceptibility analysis using back-propagation neural network model at Cameron Highland, Malaysia. *Landslides.* 7(1):13–30. doi: [10.1007/s10346-009-0183-2](https://doi.org/10.1007/s10346-009-0183-2).
- Schicker R, Moon V. 2012. Comparison of bivariate and multivariate statistical approaches in landslide susceptibility mapping at a regional scale. *Geomorphology.* 161–162:40–57. doi: [10.1016/j.geomorph.2012.03.036](https://doi.org/10.1016/j.geomorph.2012.03.036).

- Su Q, Zhang J, Zhao S, Wang L, Liu J, Guo J. 2017. Comparative assessment of three nonlinear approaches for landslide susceptibility mapping in a coal mine area. *IJGI*. 6(7):228. doi: [10.3390/ijgi6070228](https://doi.org/10.3390/ijgi6070228).
- Umar Z, Pradhan B, Ahmad A, Jebur MN, Tehrany MS. 2014. Earthquake induced landslide susceptibility mapping using an integrated ensemble frequency ratio and logistic regression models in West Sumatera Province, Indonesia. *Catena*. 118:124–135. doi: [10.1016/j.catena.2014.02.005](https://doi.org/10.1016/j.catena.2014.02.005).
- Wang C, Lin Q, Wang L, Jiang T, Su B, Wang Y, Mondal SK, Huang J, Wang Y. 2022. 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. *Nat Hazards*. 112(3):1967–1988. doi: [10.1007/s11069-022-05252-8](https://doi.org/10.1007/s11069-022-05252-8).
- Xu K, Guo Q, Li Z, Xiao J, Qin Y, Chen D, Kong C. 2015. Landslide susceptibility evaluation based on BPNN and GIS: a case of Guojiaba in the Three Gorges Reservoir Area. *Int J Geographical Inform Sci*. 29(7):1111–1124. doi: [10.1080/13658816.2014.992436](https://doi.org/10.1080/13658816.2014.992436).
- Yavuz Ozalp A, Akinci H, Zeybek M. 2023. Comparative analysis of tree-based ensemble learning algorithms for landslide susceptibility mapping: a case study in Rize, Turkey. *Water*. 15(14):2661. doi: [10.3390/w15142661](https://doi.org/10.3390/w15142661).
- Youssef AM, Pourghasemi HR, Pourtaghi ZS, Al-Katheeri MM. 2016. Erratum to: landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. *Landslides*. 13(5):1315–1318. doi: [10.1007/s10346-015-0667-1](https://doi.org/10.1007/s10346-015-0667-1).
- Zêzere JL, Pereira S, Melo R, Oliveira SC, Garcia RA. 2017. Mapping landslide susceptibility using data-driven methods. *Sci Total Environ*. 589:250–267. doi: [10.1016/j.scitotenv.2017.02.188](https://doi.org/10.1016/j.scitotenv.2017.02.188).
- Zhou C, Yin K, Cao Y, Ahmed B, Li Y, Catani F, Pourghasemi HR. 2018. Landslide susceptibility modeling applying machine learning methods: a case study from Longju in the Three Gorges Reservoir area, China. *Comput Geosci*. 112:23–37. doi: [10.1016/j.cageo.2017.11.019](https://doi.org/10.1016/j.cageo.2017.11.019).