China Geology 7 (2024) 104−115

China Geology

Journal homepage: http://chinageology.cgs.cn
https://www.sciencedirect.com/journal/china-geology

Comparative study of different machine learning models in landslide susceptibility assessment: A case study of Conghua District, Guangzhou, China

Ao Zhang[a], Xin-wen Zhao[a], Xing-yuezi Zhao[a,*], Xiao-zhan Zheng[b], Min Zeng[a], Xuan Huang[c], Pan Wu[a], Tuo Jiang[a], Shi-chang Wang[a], Jun He[a], Yi-yong Li[a]

[a] *Wuhan Center, China Geological Survey, Ministry of Natural Resources (Geosciences Innovation Center of Central South China), Wuhan* 430205, *China*
[b] *Guangzhou Institute of Geological Survey, Guangzhou* 510080, *China*
[c] *Hubei Transportation Planning Design Institute Co., Ltd, Wuhan* 430050, *China*

ARTICLE INFO

Article history:
Received 4 June 2023
Received in revised form 3 August 2023
Accepted 29 August 2023
Available online 27 September 2023

Keywords:
Landslides susceptibility assessment
Machine learning
Logistic Regression
Random Forest
Support Vector Machines
XGBoost
Assessment model
Geological disaster investigation and prevention engineering

ABSTRACT

Machine learning is currently one of the research hotspots in the field of landslide prediction. To clarify and evaluate the differences in characteristics and prediction effects of different machine learning models, Conghua District, which is the most prone to landslide disasters in Guangzhou, was selected for landslide susceptibility evaluation. The evaluation factors were selected by using correlation analysis and variance expansion factor method. Applying four machine learning methods namely Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM), and Extreme Gradient Boosting (XGB), landslide models were constructed. Comparative analysis and evaluation of the model were conducted through statistical indices and receiver operating characteristic (ROC) curves. The results showed that LR, RF, SVM, and XGB models have good predictive performance for landslide susceptibility, with the area under curve (AUC) values of 0.752, 0.965, 0.996, and 0.998, respectively. XGB model had the highest predictive ability, followed by RF model, SVM model, and LR model. The frequency ratio (FR) accuracy of LR, RF, SVM, and XGB models was 0.775, 0.842, 0.759, and 0.822, respectively. RF and XGB models were superior to LR and SVM models, indicating that the integrated algorithm has better predictive ability than a single classification algorithm in regional landslide classification problems.

1. Introduction

Landslide is one of the more frequent geological disasters in China, often causing serious damage to natural resources, the ecological environment, and infrastructure, posing a serious threat to the safety of people ’s lives and property (Guzzetti F et al., 2012). Against the backdrop of global climate change, the number of extreme weather events in China has increased. The originally fragile geological environment has exacerbated the risk of landslides, and landslide prevention and mitigation work has become an urgent task in current society. Landslide susceptibility evaluation is an important basis for landslide disaster prevention and mitigation. Landslide susceptibility refers to the possibility of landslide occurrence under certain geological and environmental conditions in a certain area, with a focus on evaluating the probability of the location and spatial aspect of landslide occurrence (Reichenbach P et al., 2018).

In the process of landslide susceptibility assessment, the most critical link is to establish an appropriate assessment model. An excellent vulnerability assessment model can fully tap the mapping relationship between landslide and its basic environmental factors, and build a nonlinear function from basic environmental factors to landslide spatial probability (Huang FM et al., 2022; Dou J et al., 2019; Tsangaratos P et al., 2017; Jia YF et al., 2023; Xiong XH et al., 2022). The commonly used landslide susceptibility evaluation models at home and abroad are divided into non-deterministic models and deterministic models (Xia H et al., 2018). In the development process of landslide susceptibility evaluation models, traditional non-deterministic models based on

First author: *E-mail address*: zhangao@mail.cgs.gov.cn (Ao Zhang).

* Corresponding author: *E-mail address*: zhaoxingyuezi@mail.cgs.gov.cn (Xing-yuezi Zhao).

Literary editor: Li-qiong Jia
doi:10.31035/cg2023056
2096-5192/© 2024 China Geology Editorial Office.

statistical analysis have been widely applied in previous studies, such as information quantity models (Liu HH, 2012; Wang T et al., 2021), evidence weight models (Li JL et al., 2016), analytic hierarchy processes (Yang DH et al., 2015), fuzzy comprehensive evaluation methods (Chen W et al., 2021), etc.

With the development of data mining and artificial intelligence, domestic and foreign scholars have gradually applied algorithm models in the field of machine learning to landslide susceptibility evaluation, such as Logical Regression (LR) (Zhang J et al., 2016), Decision Tree (DT) (Paraskevas T et al., 2016), Support Vector Machine (SVM) (Chen W et al., 2016), Artificial Neural Network (ANN) (Dou J et al., 2015), etc. Lee S et al. (2001) used LR method and probability method to generate landslide susceptibility zoning map, and the evaluation results were consistent with the landslide survey data. Feng HJ et al. (2016) compared the application of LR, Information Quantity (IQ), and ANN in landslide susceptibility evaluation in Chun 'an County, Zhejiang Province, and concluded that the ANN model was superior to the other two models. Zêzere JL et al. (2017) and Huang Y et al. (2018) carried out a comparative analysis of machine learning models to predict landslide susceptibility, and believed that the SVM model had the advantages of relatively stable prediction results and high recognition. Based on the traditional weighting system, Kanungo DP et al. (2006) used ANN model, fuzzy weighting model and a hybrid model of neural network and fuzzy weighting to evaluate landslide susceptibility. The results showed that the hybrid model of neural networks and fuzzy weighting was more accurate.

To further verify the generalization performance of machine learning models in landslide susceptibility evaluation, and promote the application of these machine learning models in the landslide field, the typical machine learning medels namely Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM), and Extreme Gradient Boosting (XGB) are selected to evaluate the landslide susceptibility. Statistical indexs and receiver operating characteristic (ROC) curve were used to verify the prediction effect of each model.

2. Overview of the study area and data sources

2.1. Overview of the study area

Conghua District of Guangzhou is located in the middle of Guangdong Province, 113°17′–114°04′E, 23°22′–23°56′N, with a total area of 1974.5 km^2, which is the highest altitude district in Guangzhou. This study area is located in the transition zone from the Pearl River Delta to the mountainous area of northern Guangdong, with the terrain inclined from north to south. The highest point is Liangkou Tiantangding, with an altitude of 1210 m, and the lowest point is Taiping Village, Taiping Town, with an altitude of 16 m (Fig. 1). The

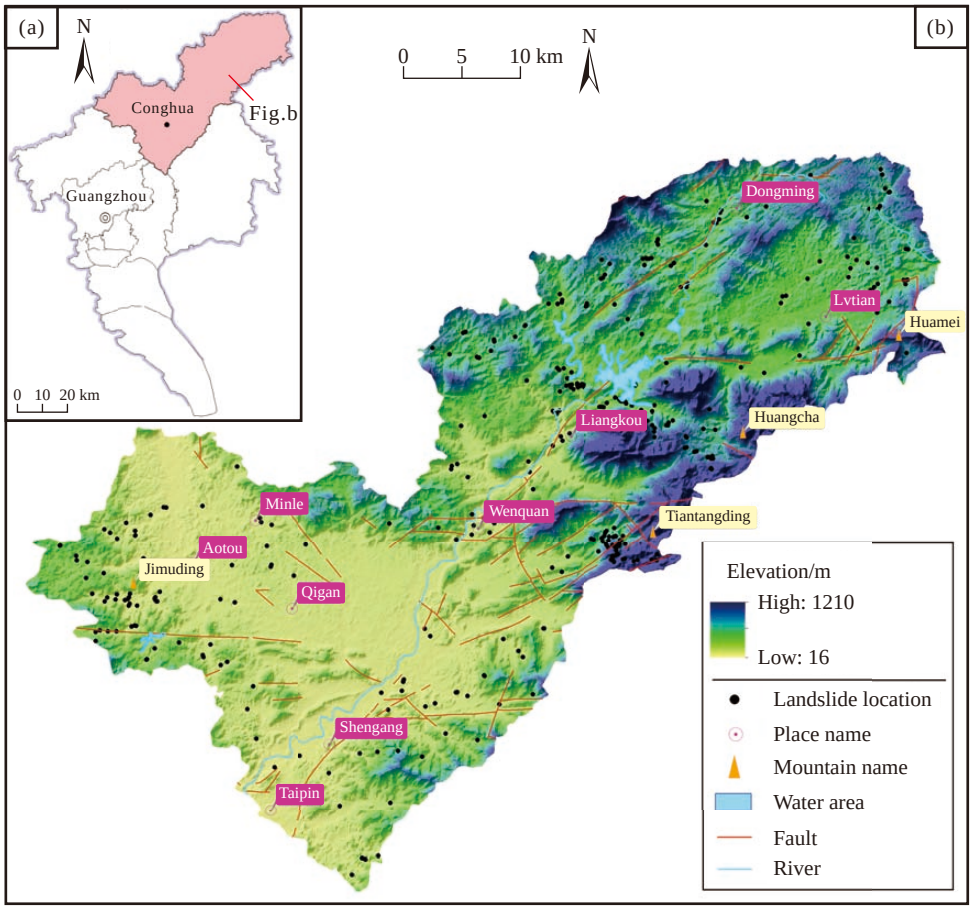

Fig. 1. Location of the study area (a) and geographical location of the study area and historical landslide distribution map (b).

landform is dominated by low mountains and hills, with steep terrain and developed valleys. The study area has a humid monsoon climate in the north subtropical zone with a mild climate and abundant rainfall. The annual average temperature is 21–22°C and the annual rainfall is 1907 mm.

The main exposed strata in the area are Proterozoic gneiss complex, Devonian siltstone, Carboniferous limestone dolomite, Triassic quartz sandstone, Jurassic– Cretaceous granite, granodiorite, Paleogene sandstone, glutenite, and Quaternary clay. The complex and changeable geological and geomorphic conditions provide favorable conditions for the breeding of landslides. According to statistics, 1231 landslides occurred in Guangzhou from 2013 to 2020. There were 365 landslides in the Conghua District, accounting for 29.7%, which was one of the areas with the largest number of landslides. Landslides were mainly distributed in middle and low mountains, hills, and river banks with strong topographic cutting in the area.

2.2. Data sources

In the study of landslide susceptibility, the relationship between historical landslide events, geological environmental factors, and location space must be considered (Pham BT et al., 2015, 2016). Therefore, the initial data in this study include: (1) The detailed survey data of geological disasters in the Conghua District, which were used to obtain the location distribution of historical landslide points. The data were derived from the "Guangzhou multi-factor urban geological survey " and "Guangzhou 1∶50000 geological disaster detailed survey " project; (2) the digital elevation model (DEM) with a spatial resolution of 30 m, which was derived from the free and public basic data of the Geographic Data Cloud (https://www.gscloud.cn) and used to analyze and extract the basic factors of topography and geomorphology, such as elevation, slope, and aspect; (3) Landsat-8 remote sensing image with a spatial resolution of 30 m, which can also be obtained from the Geographic Data Cloud and was used to obtain normalized difference vegetation index (NDVI); (4) meteorological station observation data, which were derived from National Earth System Science Data Center, National Science and Technology Infrastructure of China (http://www.geodata.cn) and were used to obtain the annual average rainfall factor. According to the size of the study area and the scale of landslides, the grid with a resolution of 30 m×30 m was used as the basic unit for landslide susceptibility assessment. The spatial analysis of all initial data was carried out to extract the landslide susceptibility evaluation factors. After rasterization, the spatial resolution of all factors was unified to 30 m.

3. Methodology

In this landslide susceptibility assessment, the prediction performance of different machine learning landslide models is compared and analyzed to draw a more accurate landslide susceptibility zoning map in the study area. The research process is mainly divided into six steps (Fig. 2). (1) Data acquisition: The geospatial database of basic environmental factors is established based on the landslide frequency ratio. (2) Feature extraction: Feature selection method is used to screen appropriate susceptibility evaluation feature indicators. (3) Data set division and Feature Engineering: The training set and test set are divided and standardized. (4) Machine learning: Using training sets to build landslide models with different algorithms. (5) Model evaluation: Use the test set to verify and compare the accuracy of different landslide models. (6) Drawing landslide susceptibility map.

3.1. Frequency ratio method

Previous studies have shown that locations with similar geological and topographical conditions as historical landslide points are more prone to landslide (Miao WD et al., 2003), and the frequency ratio (FR) method can better reflect the nonlinear relationship between the basic environment and landslide susceptibility, which has been widely used in the field of landslide (Chen W et al., 2021; Hu T et al., 2020). Therefore, in this paper, the FR method based on the principle of statistics was used as a quantitative model to calculate the ratio of the frequency of known landslide units in each partition to the frequency of all grid units in the corresponding partition in the study area, so as to quantify the contribution of each factor interval grade to landslide susceptibility (Equation 1).

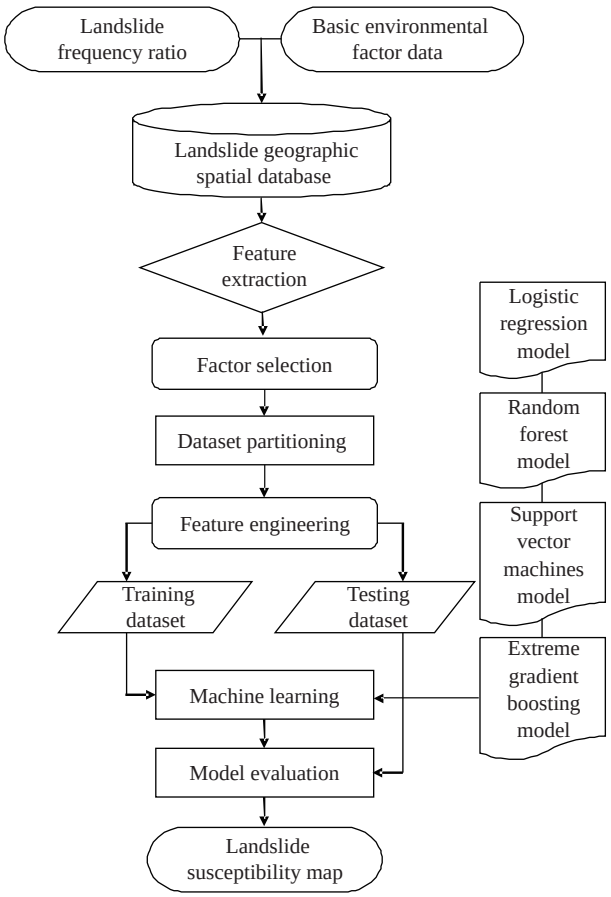

Fig. 2. Flowchart of the machine learning model for landslide susceptibility evaluation.

$$FR = \frac{n_i/N}{s_i/S} \quad (1)$$

Where n_i is the number of landslide grids in the i-th grading interval for each type of factor; N is the total number of landslide grids in the area; s_i is the grid number of each type of factor in the i-th hierarchical interval; S is the total number of grids in the study area; FR is the frequency ratio of these factors.

3.2. Landslide Susceptibility Prediction Model

3.2.1. Logistic Regression (LR)

Landslide susceptibility assessment is to determine whether a unit is a landslide one. There are only two answers: "yes" and "no", so it is a binary problem. As a typical classification model in machine learning, LR has the advantages of simple and efficient algorithm, and is a powerful tool to solve the problem of binary classification. Based on the core mathematical concept of natural logarithm (Peng CYJ et al., 2002), LR was introduced and widely used in the late 1960s and early 1970s (Cabrera AF, 1994), which is very suitable for describing and verifying the nonlinear relationship between the classification result variable (landslide or non landslide) and the classification prediction variable (landslide impact factor). The natural logarithm of LR is expressed as Equation 2:

$$Y = f(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad (2)$$

Therefore, the probability of landslide event (P) can be determined by Equation 3:

$$P = P(Y|X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}} \quad (3)$$

Where Y is the result variable (landslide or non landslide); $X = X_1, X_2, \cdots X_n$ represent environmental factors affecting landslide; n is the number of environmental factors; β_0 is the intercept; $\beta_1, \beta_2, \cdots, \beta_n$ is the regression coefficient.

3.2.2. Random Forest (RF)

RF is a typical representative of ensemble learning. Combining Breiman's idea and Ho's description method, multiple decision trees are constructed through different data subsets, and then the judgment results of multiple decision trees are voted to obtain the final output of a random forest. Compared with the traditional landslide evaluation method, the new method of random sampling of samples and characteristics is introduced, which can improve the accuracy and stability of the model more than a single decision tree. A large number of studies have shown that random forest has a high fault tolerance rate in terms of algorithm, outliers, and noise.

RF adopts a bagging algorithm. Samples are randomly selected to construct a subset of data and have a sample to be put back. m feature attributes are randomly selected from all feature attributes M to build a weak decision tree (m<M). The machine training n weak decision trees $y_1(X), y_2(X), \cdots y_i(X)$ are obtained by repeating n times, and the random forest model is established. Its expression is (Equation 4):

$$Y(x) = arg_Z^{max} \sum_{i=1}^{n} I(y_i(X) = Z) \quad (4)$$

Where $Y(x)$ is the RF model; $y_i(X)$ is the single weak decision tree model; Z is the output variable; I is the explicit function; n is the number of weak decision trees.

3.2.3. Support Vector Machines (SVM)

SVM is a machine learning method based on structural risk minimization principle. Through non-linear mapping, the non-linear separable data is mapped to the high-dimensional feature space, and the optimal classification hyperplane is found in this feature space to realize the efficient classification of positive and negative data. It can keep the interval to the maximum, which makes the support vector machine have better robustness. The hyperplane calculation formula is (Equation 5):

$$w^{\mathrm{T}}x + b = 0 \quad (5)$$

Where w is the normal vector; x is the feature vector of sample points; b is a constant. When w and b are optimal, it means that the optimal hyperplane is found so that the distance between positive and negative samples is the largest. By solving the above optimization problem and introducing relaxation variables and penalty factors in the calculation process, the optimal hyperplane can be determined as (Equations 6, 7):

$$\min\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i = 0 \quad (6)$$

$$\text{s.t.} y_i\left(w^{\mathrm{T}}x_i + b\right) \geqslant 1 - \xi_i \quad (7)$$

Lagrange multipliers are introduced to transform the dual problem, and the optimal classification decision function is obtained (Equation 8).

$$f(x) = sign\left[\sum_{i=1}^{n}\alpha_i y_i \kappa\left(x_i, x_j\right) + b\right] \quad (8)$$

Where $\kappa\left(x_i, x_j\right) = \varphi(x_i)\cdot\varphi\left(x_j\right)$ is the kernel function, representing the mapping from the input space to the feature space. Linear kernel (LN), polynomial kernel (PL), radial basis function (RBF) and sigmoid kernel (SIG) are commonly used in SVM.

3.2.4. Extreme Gradient Boosting (XGB)

XGB, as one of the newly proposed algorithms, is the trump card of integrated learning. At present, it is rarely used in landslide susceptibility evaluation. Compared with the traditional gradient lifting tree, XGB is faster than other integration algorithms using gradient lifting and has been considered an advanced evaluator with ultra-high performance in classification and regression. XGB algorithm changes to a greedy strategy for the continuous iteration of weight distribution in the classification of wrong samples. The best direction of training is the direction of loss function

gradient decline. Taylor's second-order expansion is used to optimize the loss function. At the same time, a regularization term is added to control the complexity of the model to prevent overfitting.

XGB builds the optimal model by minimizing the loss function. The loss function is a regular term that increases the complexity of the model. The objective function is set to optimize the iterative model using the idea of minimizing structural risk. The objective function of the model is expressed as Equation 9:

$$obj^{(t)} = \sum_i^n L\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^t \Omega(f_k) \tag{9}$$

Where $\hat{y}_i$ is the model prediction value of the i-th sample in the t-round; y_i is the real value; $L\left(y_i, \hat{y}_i^{(t)}\right)$ is the prediction error of the t-round; n is the total number of samples; $\Omega(f_k)$ is the regularization term of the k -round, representing the complexity of the k-round model, defined as Equation 10:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{10}$$

Where T is the number of leaf nodes; $\|w\|$ is the modulus of leaf node vector; γ is the difficulty of node segmentation; λ is the regularization coefficient.

The objective function is optimized by Taylor second-order expansion. Ordering $g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$, $h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$, bring Equation 10 into Equation 9, Equation 9 is expressed as Equation 11:

$$obj^{(t)} = \sum_{j=1}^T \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T \tag{11}$$

Where $I_j = \{i | q(x_i) = j\}$, is the sample set of the jth leaf node; $q(x_i)$ is the structure of the tree. Ordering $\sum_{i \in I_j} g_i = G_j$, $\sum_{i \in I_j} h_i = H_j$, G_j and H_j are deterministic quantities, and Equation 11 can be regarded as a quadratic function of one variable with respect to leaf node w. Minimizing Equation 11, the optimal parameters and the optimal loss function are obtained (Equations 12, 13):

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{12}$$

$$obj^* = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{13}$$

3.3. Evaluation and comparison methods

To verify the performance and generalization ability of the model, it is necessary to evaluate the fitness of the model in different data sets (Tien Bui D et al., 2016). The evaluation results in the training data set reflect the fitting degree of the landslide model and training data. The evaluation results in the test data set reflect the prediction ability of the landslide model (Tien Bui D et al., 2012). In this paper, the performance of four landslide models is evaluated and compared by using the evaluation based on statistical indicators and the ROC curve.

3.3.1. Evaluation based on statistical index

Many indexes in statistics can be used to verify the performance of the model (Bennett ND et al., 2013). In this paper, positive predictive value, negative predictive value, sensitivity, specificity, accuracy, root mean square error, and other statistical indicators are used to verify the performance of the landslide model (Table 1).

Where TP (true positive) is the number of landslide points correctly classified as landslide; TN (true negative) is the total number of non-landslide points correctly classified as non-landslide; FN (false negative) is the number of landslides classified as non-landslide; FP (false positive) is the number of non-landslide points classified as a landslide. e_i is the estimated value of the i-th observation, $\overline{e_i}$ is the measured value of the i-th observation.

3.3.2. Receiver operating characteristic curve

The receiver operating characteristic (ROC) curve is widely used in model comparison and evaluation. ROC curve is generated by counting the sensitivity (landslide samples predicted as a landslide) and 1-specificity (non-landslide samples predicted as a landslide) of each model. The area under curve (AUC) value of the ROC curve is between 0 and 1. The larger the value, the higher the prediction accuracy. When AUC is equal to 1, the model is a perfect classifier.

4. Machine learning and result analysis

4.1. Evaluation factors and analysis

4.1.1. Selection and quantification of basic environmental factors

In the study of landslide susceptibility evaluation, the selection of appropriate evaluation factors directly affects the

Table 1. Description of statistical index.

No.	Name	Formula	Description
1	Sensitivity (SST)	$SST = \frac{TP}{TP+FN}$	*SST* represents the percentage of landslide grid correctly classified as "landslide" predicted
2	Specificity(SPF)	$SPF = \frac{TN}{FP+TN}$	*SPF* represents the percentage of non-landslide grids correctly classified as "non-landslide"
3	Accuracy(ACC)	$ACC = \frac{TP+TN}{TP+TN+FP+FN}$	*ACC* represents the proportion of "landslide" and "non-landslide" correctly classified in the total grid
4	Root mean squared error (RMSE)	$RMSE = \left[(1/m)\sum_{i=1}^m (e_i - \overline{e_i})^2\right]^{0.5}$	*RMSE* shows the error metric in the same units with the original data Smaller RMSE value indicates better performance of landslide model

reliability and accuracy of susceptibility results. Selection is generally based on the objective existence, significance, and inheritance of evaluation factors (Reichenbach P et al., 2018; Huang FM et al., 2020). According to the formation conditions of landslide disaster in the study area, landslide mainly depends on the combined action of internal and external factors of geological environment conditions. In this study, a total of seven indicators including elevation, gradient, slope aspect, curvature, annual rainfall, NDVI, and stratum lithologic were selected as the preliminary basic environmental factors. These evaluation factors have been applied in many studies as the main factors affecting the formation of landslides, and can easily obtain high-precision data through investigation and collection to ensure the accuracy of landslide susceptibility evaluation. The natural discontinuity method was used to divide the selected basic environmental factors into five attribute intervals. The division results are shown in Fig. 3. The FR of each environmental factor attribute interval was calculated according to Equation 1, and the calculation results are shown in Table 2. The FR was used as the quantitative value of each factor for machine learning. The size of the FR determines whether the attribute interval of environmental factors is conducive to landslide development.

4.1.2. Feature analysis

The basic environmental factors selected are related to the occurrence of landslides to a certain extent, but the possible correlation between the factors will adversely affect the prediction results and may increase the complexity and running time of the model. Therefore, before the evaluation, it is necessary to select the features of the quantified factors and carry out correlation analysis to eliminate the factors with high correlation to improve the model efficiency and prediction accuracy. The correlation analysis results (Fig. 4) show that except for each factor's autocorrelation coefficient of 1, the correlation coefficients of the other factors are small, which are −0.22–0.23, indicating that each basic environmental factor can be independently used as the characteristic variable of landslide susceptibility assessment.

Because the collinearity between sample data easily leads to the decline of model training accuracy, the data set used to build the model needs to meet certain collinearity requirements. Variance in inflation factor (VIF) is used to characterize the degree of collinearity between factors. When VIF is greater than 10, it indicates that there is serious collinearity between data and it needs to be eliminated. The analysis results are shown in Table 3. The VIF value of each basic environmental factor does not exceed 10, indicating that the sample data is not significantly collinear, and all basic environmental factors can be used as evaluation factors to evaluate landslide susceptibility without elimination.

4.2. Landslide susceptibility model training, prediction and evaluation

The natural breakpoint method and FR method were used to quantify the selected seven types of evaluation factors, and the geospatial database of landslide susceptibility evaluation in the study area was obtained. To build the landslide prediction model, the database needs to be extracted and segmented, and the training set and test set are generated respectively. The training data set was used to train the landslide model, and the test data set was used to verify the performance of the landslide model (Pham BT et al., 2016).

All 365 landslide points in the study area were taken as landslide sample points, and 1000 non-landslide sample points were randomly extracted outside the landslide area. The landslide sample points and non-landslide sample points were combined to form a sample data set. 70% (955 points) were randomly selected as the training set, and the remaining 30% (410 points) were selected as the test set. 30 m×30 m grid as the basic evaluation unit, the study area is divided into 2209297 units. LR, RF, SVM, and XGB landslide models were constructed using the training set, and the performance of the four landslide models on the training set and the test set was analyzed respectively (Tables 4, 5). The results indicate that all landslide models show high prediction ability. In the training set, XGB model (ACC=0.99) has the highest ACC value, followed by RF model (ACC=0.94), SVM model (ACC=0.82), and LR model (ACC=0.80); In the test set, XGB model (ACC=0.99) has the highest ACC value, followed by RF model (ACC=0.97), SVM model (ACC=0.83) and LR model (ACC=0.83). Therefore, XGB model has the highest prediction ability, followed by RF, SVM, and LR models.

In addition, the ROC curve using the training set is shown in Fig. 5. The AUC values of SVM, RF, and XGB models are relatively close, which are 0.999, 0.997, and 0.991 respectively. The AUC value of LR model is the lowest (AUC=0.722); The ROC curve using the test set is shown in Fig. 6. The AUC values of RF and XGB models are relatively close, which are 0.998 and 0.996 respectively. The AUC value of SVM model is slightly smaller (AUC=0.965), and the AUC value of LR model is the lowest (AUC=0.752). The results of ROC curve analysis show that the prediction ability of LR model is the worst among the four models, the prediction ability of SVM, RF, and XGB models is similar, and the prediction ability of XGB model is slightly stronger.

4.3. Validation and Comparison of Landslide Susceptibility Models

Four landslide models trained and tested were used to calculate the landslide susceptibility index from all grid data in the study area. The natural breakpoint method was used to classify the landslide susceptibility index into five grades: Very low, low, moderate, high, and very high. The landslide susceptibility map (Fig. 7) predicted by each model was compiled. The results show that the prediction results of all models are similar in regional spatial distribution, reflecting the characteristics that the distribution law of landslide susceptibility is consistent with the basic geographical

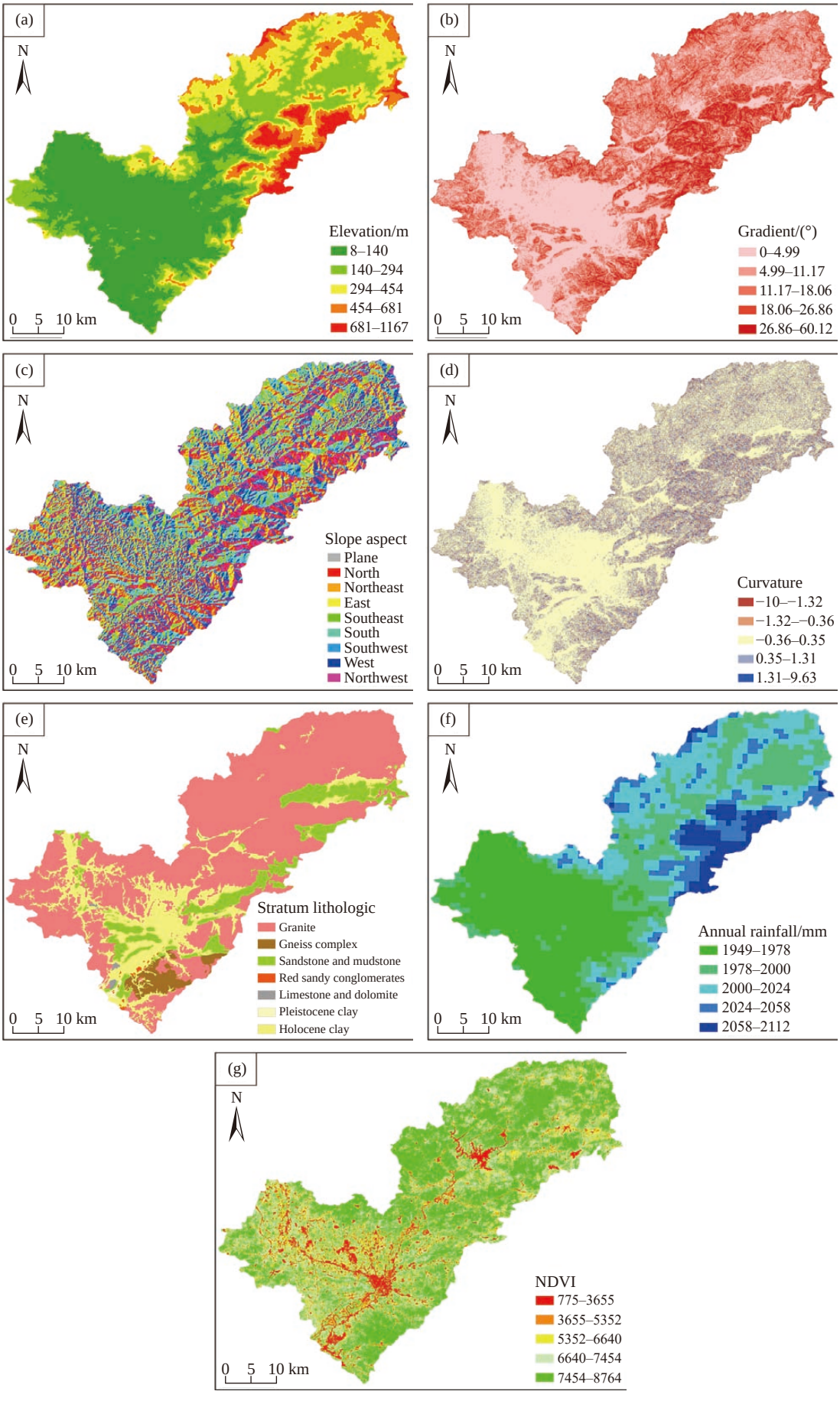

Fig. 3. Basic environmental factors of landslide susceptibility assessment.

Table 2. Attribute interval and FR of each evaluation factor.

Factors	Attribute interval	Number of interval landslide grids	Interval grid number	FR
Elevation/m	8–139	120	914164	0.795
	139–294	114	529250	1.304
	294–453	79	434705	1.100
	453–681	32	230855	0.839
	681–1167	20	100323	1.207
Gradient/°	0–4.99	114	765679	0.901
	4.99–11.17	128	585808	1.323
	11.17–18.06	79	454046	1.053
	18.06–26.86	34	279536	0.736
	26.86–60.12	10	124228	0.487
Slope aspect	Plane	1	5544	1.092
	North	60	297253	1.222
	Northeast	30	215802	0.841
	East	48	252204	1.152
	Southeast	49	277307	1.070
	South	59	333308	1.071
	Southwest	52	251346	1.252
	West	31	281488	0.667
	Northwest	35	295045	0.718
Curvature	−10.7–−1.32	5	72633	0.417
	−1.32–−0.36	57	304845	1.132
	−0.36–0.35	222	1230177	1.092
	0.35–1.31	66	484044	0.825
	1.31–9.63	15	117598	0.772
Annual rainfall/mm	1949–1978	87	792643	0.664
	1978–2000	101	548694	1.114
	2000–2024	88	503742	1.057
	2024–2058	54	232038	1.409
	2058–2112	35	132180	1.603
NDVI	775–3655	30	107838	1.684
	3655–5352	66	127002	3.146
	5352–6640	67	255380	1.588
	6640–7454	111	699936	0.960
	7454–8764	91	1019141	0.540
Stratum lithology	Granite	252	1490524	1.023
	Gneiss complex	3	58843	0.309
	Sandstone and mudstone	58	247037	1.421
	Red sandy conglomerates	0	3338	0.000
	limestone and dolomite	1	5379	1.125
	Pleistocene clay	0	5506	0.000
	Holocene clay	51	398670	0.774

environment. The very low and low susceptible areas of landslide in the Conghua district are concentrated in the alluvial pluvial plain in the southeast, while the very high and high susceptible areas are mainly concentrated in the valleys in the northeast and low mountain areas in the East, extending radially to the northwest, north, northeast, east and other directions. The landslide is mainly affected by the steep terrain in the northeast of Conghua District, the thick and loose residual overburden in the valley, rainfall scouring, slope cutting, and other human engineering activities. The historical landslide points are mostly located in the middle and high prone areas (Fig. 8). The predicted results are relatively consistent with the distribution of historical landslide disaster points.

Fig. 9 shows the percentage of grid number (or area) in the very low, low, medium, high, and very high regions of all models. Statistics shows that the susceptibility of each model presents a similar distribution law. The susceptibility of the study area is mainly moderate, followed by high, low, and very low levels, and the lowest high level. LR, RF, SVM, and XGB models account for 12.71%, 8.48%, 12.89%, and 8.43%, respectively.

The actual prediction accuracy of each model was compared and analyzed by using the FR method, and the FR of each landslide model was calculated according to Equation 1 (Table 6). The FR accuracy of landslide susceptibility results can be obtained by dividing the FR of very high and high-susceptible areas by the sum of all FR (Huang FM et al., 2022). The FR of LR model prediction results from very low to very high prone areas is 0.227, 0.405, 0.650, 1.405, 3.018, and its FR accuracy is 0.775. The FR of the prediction results of RF model from very low to very high prone areas is 0.173, 0.344, 0.550, 1.431, 4.263, and the accuracy of FR is 0.842. The FR of the prediction results of SVM model from very low to very high prone areas is 0.315, 0.505, 0.602, 1.161, 3.316, and its FR accuracy is 0.759. The FR of the prediction results of XGB model from very low to very high prone areas is 0.185, 0.355, 0.720, 1.562, 4.256, and its FR accuracy is 0.822. The comparison results show that RF and XGB models are superior to LR and SVM models in terms of FR accuracy. RF and XGB models are typical representatives of the integrated algorithm. It can be seen that the integrated algorithm has more advantages in classification, and the predicted landslide susceptibility can better reflect the spatial distribution law and environmental aggregation characteristics of landslides.

5. Conclusions

Taking the Conghua District of Guangzhou as the study area, the correlation analysis and variance expansion factor method were used to select the basic environmental factors such as elevation, gradient, slope aspect, curvature, annual rainfall, NDVI, stratum lithologic, and the landslide susceptibility in the study area was predicted and evaluated based on LR, RF, SVM, XGB models.

(i) Statistical index analysis shows that the four models have good prediction performance of landslide susceptibility. In the test set, the ACC values of LR, RF, SVM, and XGB models are 0.83, 0.97, 0.83, and 0.99 respectively, and the AUC values of ROC curves are 0.752, 0.965, 0.996, and 0.998 respectively. XGB model has the highest prediction ability, followed by RF, SVM, and LR models.

(ii) Based on the four landslide models, landslide susceptibility maps were compiled, showing similar spatial distribution characteristics. The susceptibility of the Conghua district is mainly moderate, followed by high, low, and very

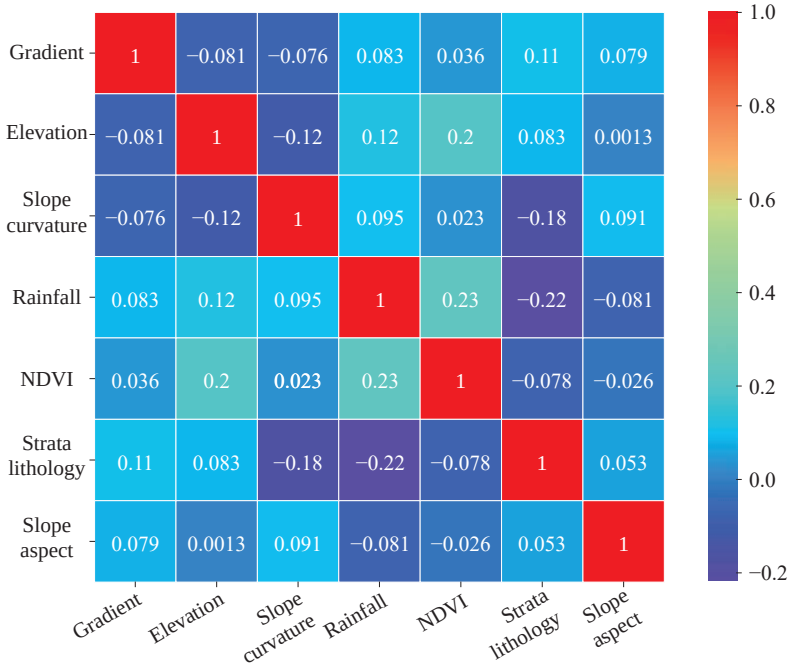

Fig. 4. Correlation analysis of evaluation factors of heat map.

Table 3. Collinearity analysis of basic environmental factors.

No.	Factors	VIF
1	Gradient	1.066933
2	Elevation	1.336441
3	Curvature	1.072272
4	Annual rainfall	1.144736
5	NDVI	1.255478
6	Stratum lithologic	1.113084
7	Slope aspect	1.031587

Table 4. Model performance using training dataset.

No.	Parameters	LR	SVM	RF	XGB
1	TN	693	700	700	695
2	FP	7	0	0	5
3	FN	181	169	54	4
4	TP	74	86	201	251
5	SST/%	0.29	0.34	0.79	0.98
6	SPF/%	0.99	1.00	1.00	0.99
7	ACC	0.80	0.82	0.94	0.99
8	RMSE	0.44	0.42	0.24	0.10

Table 5. Model predictive capability using testing dataset.

No.	Parameters	LR	SVM	RF	XGB
1	TN	294	300	300	296
2	FP	6	0	0	4
3	FN	67	69	14	1
4	TP	63	41	96	109
5	SST/%	0.48	0.37	0.87	0.99
6	SPF/%	0.98	1.00	1.00	0.99
7	ACC	0.83	0.83	0.97	0.99
8	RMSE	0.42	0.41	0.18	0.11

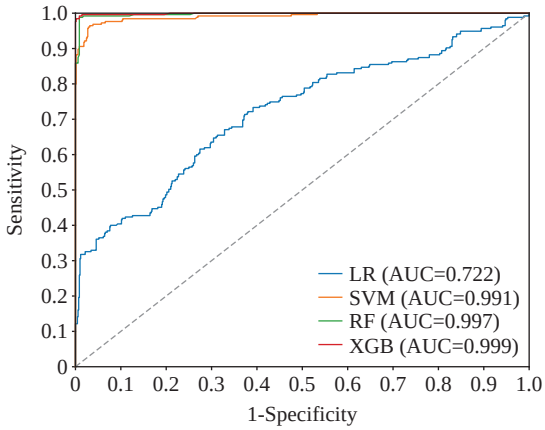

Fig. 5. Analysis of the ROC curve of different landslide models using training dataset.

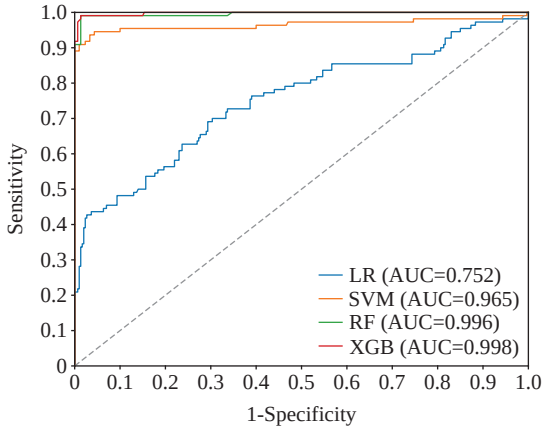

Fig. 6. Analysis of the ROC curve of different landslide models using testing dataset.

low grades, and the very high grade is the least. The proportions of very high susceptible areas predicted by LR, RF, SVM, and XGB models are 12.71%, 8.48%, 12.89%, and 8.43% respectively. The very low and low susceptible areas in

the Conghua district are concentrated in the alluvial pluvial plain in the southeast, while the extremely high and high susceptible areas are distributed in the valleys in the northeast and low mountain areas in the east, and the predicted results areconsistent with the distribution of historical landslide disaster points.

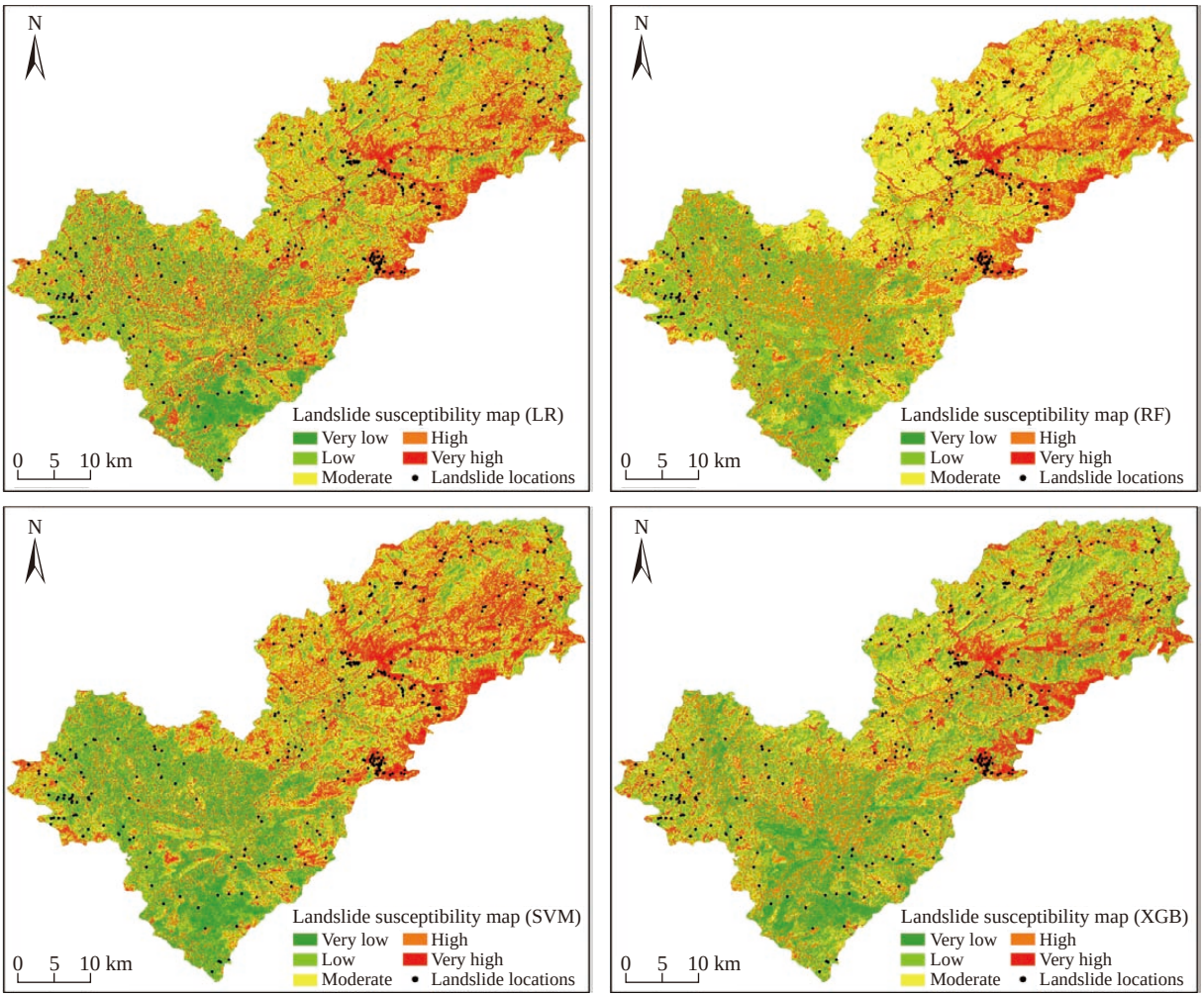

Fig. 7. Landslide susceptibility maps using different landslide models.

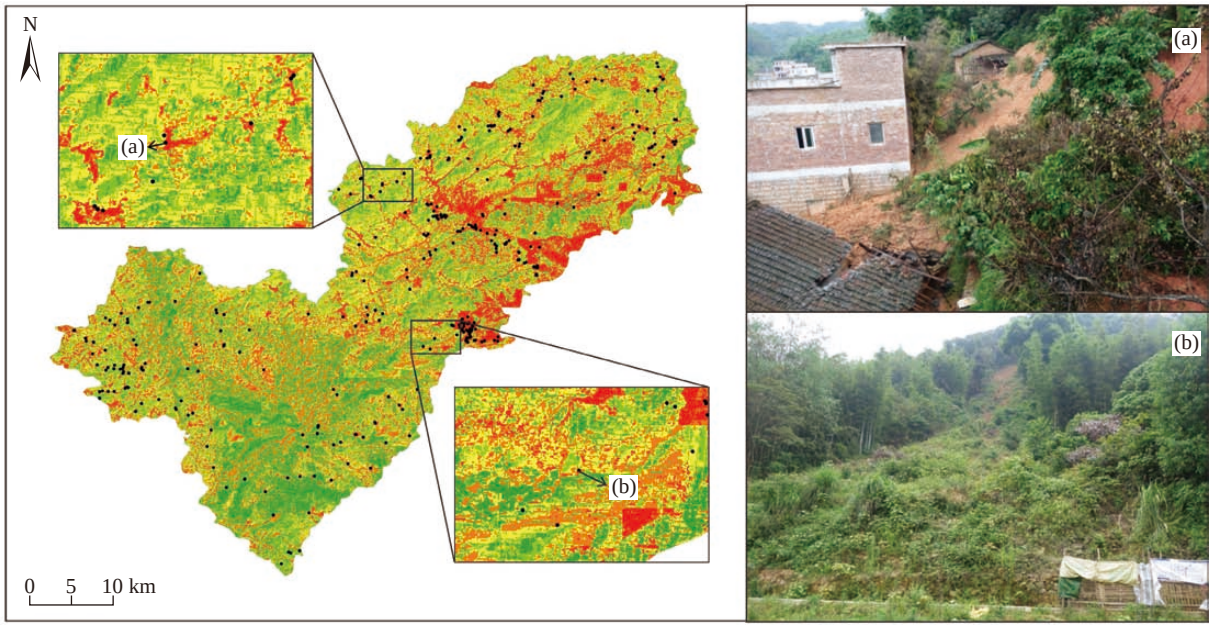

Fig. 8. Typical landslides verification.

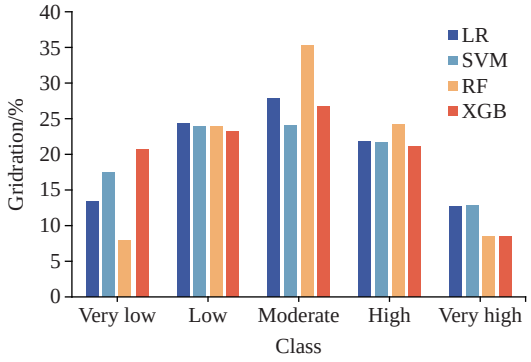

Fig. 9. Different classes distribution of grid ratio on slope map.

Table 6. FR precision analysis of susceptibility maps of different models.

Model	Class	Number of landslides	Landslide ratio/%	Number of grids	Grid ratio/%	FR
LR	Very low	11	3.01	293269	13.27	0.227
	Low	36	9.86	537750	24.34	0.405
	Moderate	66	18.08	615047	27.84	0.650
	High	112	30.68	482443	21.84	1.405
	Very high	140	38.36	280788	12.71	3.018
RF	Very low	5	1.37	174927	7.92	0.173
	Low	30	8.22	528359	23.92	0.344
	Moderate	71	19.45	781376	35.37	0.550
	High	127	34.79	537195	24.32	1.431
	Very high	132	36.16	187440	8.48	4.263
SVM	Very low	20	5.48	384467	17.40	0.315
	Low	44	12.05	527571	23.88	0.505
	Moderate	53	14.52	532830	24.12	0.602
	High	92	25.21	479657	21.71	1.161
	Very high	156	42.74	284772	12.89	3.316
XGB	Very low	14	3.84	458796	20.77	0.185
	Low	30	8.22	511172	23.14	0.355
	Moderate	70	19.18	588132	26.62	0.720
	High	120	32.88	464890	21.04	1.562
	Very high	131	35.89	186307	8.43	4.256

(iii) Combining with the FR method, the prediction accuracy of the four landslide models in the whole study area was compared and analyzed. The FR accuracy of LR, RF, SVM, and XGB models are 0.775, 0.842, 0.759, and 0.822 respectively. RF and XGB models are superior to LR and SVM models, indicating that the integrated algorithm has better prediction ability in the regional landslide classification problem. The integrated algorithm based on the combination of multiple weak models is superior to the single classification algorithm in solving the single prediction problem. It can better reflect the spatial distribution law and environmental accumulation characteristics of landslides.

CRediT authorship contribution statement

Ao Zhang, Xing-yuezi Zhao, Xin-wen Zhao and Xiao-zhan Zheng conceived of the presented idea. Ao Zhang, Xin-wen Zhao and Xing-yuezi Zhao carried out the experiment. All authors discussed the results and contributed to the final manuscript.

Declaration of competing interest

The authors declare no conflicts of interest.

Acknowledgment

This research was supported by the projects of the China Geological Survey (DD20221729, DD20190291) and Zhuhai Urban Geological Survey (including informatization) (MZCD–2201–008). The authors are indebted to Guangzhou Municipal Bureau of Planning and Resources, Guangzhou Institute of Geological Survey, Guangzhou Urban Planning Survey and Design Institute for their assistance. The authors are also thankful to the reviewers and editors for their valuable comments and suggestions.

References

Bennett ND, Croke BF, Guariso G, Guillaume JH, Hamilton SH, Jakeman AJ, Marsili-Libelli S, Newham LT, Norton JP, Perrin C, Pierce SA, Robson B, Seppelt R, Voinov AA, Fath BD, Andreassian V. 2013. Characterising performance of environmental models. Environmental modelling & software, 40, 1–20 . doi: 10.1016/j.envsoft.2012.09.011.

Cabrera AF. 1994. Logistic regression analysis in higher education: An applied perspective. In: Higher Education: Handbook of Theory and Research, 10, 225–256.

Chen W, Chai HC, Zhao Z, Wang Q, Hong H. 2016. Landslide susceptibility mapping based on GIS andsupport vector machine models for the Qianyang County, China. Environmental Earth Sciences, 75(6), 1–13. doi: 10.1007/s12665-015-5093-0.

Chen W, Chen X, Peng J B, Panahi M, Lee S. 2021. Landslide susceptibility modeling based on ANFIS with teaching-learning-based optimization and satin bowerbird optimizer. Geoscience Frontiers, 12(1), 93–107. doi: 10.1016/j.gsf.2020.07.012.

Dou J, Yamagishi H, Pourghasemi HR, Yunus AP, Song X, Xu Y, Zhu Z. 2015. An integrated artificial neural network model for the landslide susceptibility assessment of Osado Island, Japan. Natural Hazards, 78(3), 1749–1776. doi: 10.1007/s11069-015-1799-2.

Dou J, Yunus A P, Bui D T, Merghadi A, Sahana M, Zhu ZF, Chen, CW, Khosravi K, Yang Y, Pham BT. 2019. Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Is-land, Japan. Science of the Total Environment, 662, 332–346 . doi: 10.1016/j.scitotenv.2019.01.221.

Feng HJ, Zhou AG, Yu JY, Tang XM, Zheng JL, Chen XX, You SY. 2016. A comparative study on Plum-Rain-Triggered landslide susceptibility assessment models in West Zhejiang Province. Earth Science, 41(3), 403–415 (in Chinese with English abstract). doi: 10.3799/dgkx.2016.032.

Guzzetti F, Mondini AC, Cardinali M, Fiorucci F, Santangelo M, Chang KT. 2012. Landslide inventory maps: New tools for an old problem. Earth-Science Reviews, 112(1–2), 42–66 . doi: 10.1016/j.earscirev.2012.02.001.

Hu T, Fan X, Wang S, Guo ZZ, Liu AC, Huang FM. 2020. Landslide susceptibility evaluation of Sinan County using logistics regression model and 3S technology. Bulletin of Geological Science and Technology, 39(2), 113–121 (in Chinese with English abstract). doi: 10.19509/j.cnki.dzkq.2020.0212.

Huang FM, Chen JW, Du Z, Yao C, Huang JS, Jiang QH, Chang ZL, Li S. 2020. Landslide susceptibility pre-diction considering regional soil

erosion based on machine-learning models. ISPRS International Journal of Geo-Information, 9(6), 377. doi: 10.3390/ijgi9060377.

Huang FM, Hu SY, Yan XY, Li M, Wang JY, Li WB, Guo ZZ, Fan WY. 2022. Landslide susceptibility prediction and identification of its main environmental factors based on machine learning models. Bulletin of Geological Science and Technology, 41(2), 79–90 (in Chinese with English abstract). doi: 10.19509/j.cnki.dzkq.2021.0087.

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.

Jia YF, Wei WH, Chen W, Yang QZ, Sheng YF, Xu GL. 2023. Landslide susceptibility assessment based on the SOM-I-SVM model. Hydrogeology & Engineering Geology, 50(3), 125–137. doi: 10.16030/j.cnki.issn.1000-3665.202206041.

Kanungo DP, Arora MK, Sarkar S, Gupta RP. 2006. A comparative study of conventional, ANN black box, fuzzy and combined neural and fuzzy weighting procedures for landslide susceptibility zonation in Darjeeling Himalayas. Engineering geology, 85(3–4), 347–366. doi: 10.1016/j.enggeo.2006.03.004.

Lee S, Min K. 2001. Statistical analysis of landslide susceptibility at Yongin, Korea. Environmental geology, 40(9), 1095–1113 . doi: 10.1007/s002540100310.

Li JL, Ma DH, Wang W. 2016. Assessment of potential seismic landslide hazard based on evidence theory and entropy weight grey incidence. Journal of Central South University (Science and Technology), 47(5), 1730–1736 (in Chinese with English abstract). doi: 10.11817/j.issn.1672-7207.2016.05.036.

Liu HH. 2012. The assessment of geohazard danger in Wenchuan County based on RS and GIS. Geology in China, 39(1), 243–251 (in Chinese with English abstract).

Miao WD. 2003. Time prediction study on occurring of landslides in Bailuyuan, Xi'an. Northwestern Geology, 36(4), 90–95 (in Chinese with English abstract).

Paraskevas T, Ioanna I. 2016. Landslide susceptibility mapping using a modified decision tree classifier in the Xanthi Perfection, Greece. Landslides, 13(2), 305–320. doi: 10.1007/s10346-015-0565-6.

Peng CYJ, Lee K, Ingersoll GM. 2002. An introduction to logistic regression analysis and reporting. The Journal of Educational Research, 96(1), 3–14. doi: 10.1080/00220670209598786.

Pham BT, Tien Bui D, Dholakia MB, Prakash I, Pham, HV. 2016. A comparative study of least square support vector machines and multiclass alternating decision trees for spatial prediction of rainfall-induced landslides in a tropical cyclones area. Geotechnical and Geological Engineering, 34, 1807–1824 . doi: 10.1007/s10706-016-9990-0.

Pham BT, Tien Bui D, Indra P, Dholakia M. 2015. Landslide susceptibility assessment at a part of Uttarakhand Himalaya, India using GIS-based statistical approach of frequency ratio method. nternational Journal of Engineering Research and Technology, 4(11), 338–344. doi: 10.17577/IJERTV4IS110285.

Reichenbach P, Rossi M, Malamud BD, Mihir M, Guzzetti F. 2018. A review of statistically-based landslide susceptibility models. Earth-Science Reviews, 180, 60–91. doi: 10.1016/j.earscirev.2018.03.001.

Tien Bui D, Pradhan B, Lofman O, Revhaug I. 2012. Landslide susceptibility assessment in Vietnam using support vector machines, decision tree, and Naive Bayes models. Mathematical Problems in Engineering, 1–26. doi: 10.1155/2012/974638.

Tien Bui D, Tuan TA, Klempe H, Pradhan B, Revhaug I. 2016. Spatial prediction models for shallow landslide hazards: A comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. Landslides, 13, 361–378. doi: 10.1007/s10346-015-0557-6.

Tsangaratos P, Ilia I, Hong H, Chen W, Xu C. 2017. Applying information theory and GIS-based quantitative methods to produce landslide susceptibility maps in Nancheng County, China. Landslides, 14(3), 1091–1111. doi: 10.1007/s10346-016-0769-4.

Wang T, Liu JM, Li ZT, Xin P, Shi JS, Wu SR. 2021. Seismic landslide hazard assessment of China and its impact on national territory spatial planning. Geology in China, 48(1), 22–39 (in Chinese with English abstract). doi: 10.12029/gc20210102.

Xia H, Yin KL, Liang X, Ma F. 2018. Landslide susceptibility assessment based on SVM-ANN Models: A case study for Wushan County in the Three Gorges Reservoir. The Chinese Journal of Geological Hazard and Control, 29(5), 13–19 (in Chinese with English abstract). doi: 10.16031/j.cnki.issn.1003-8035.2018.05.03.

Xiong XH, Wang CL, Bai YJ, Tie YB, Gao YC, Li GH. 2022. Comparison of landslide susceptibility assessment based on multiple hybrid models at county level: A case study for Puge County, Sichuan Province. The Chinese Journal of Geological Hazard and Control, 33(4), 114–124 . doi: 10.16031/j.cnki.issn.1003-8035.2022 02052.

Yang DH, Fan W. 2015. Zoning of probable occurrence level of geological disasters based on ArcGIS———A case of Xunyang. The Chinese Journal of Geological Hazard and Control, 26(4), 82–86, 93 (in Chinese with English abstract). doi: 10.16031/j.cnki.issn.1003-8035.2015.04.14.

Zêzere JL, Pereira S, Melo R, Oliveira SC, Garcia RAC. 2017. Mapping landslide susceptibility using data-driven methods. Science of the Total Environment, 589, 250–267 . doi: 10.1016/j.scitotenv.2017.02.188.

Zhang J, Yin KL, Wang JJ, Liu L, Huang FM. 2016. Study on landslide susceptibility evaluation for Wanzhou district of Three Gorges Reservoir. Chinese Journal of Rock Mechanics and Engineering, 35(2), 284–296 (in Chinese with English abstract). doi: 10.13722/j.cnki.jrme.2015.0318.