China Geology 6 (2023) 409–419

China Geology

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

Landslide susceptibility assessment in Western Henan Province based on a comparison of conventional and ensemble machine learning

Wen-geng Cao[a, b], Yu Fu[c], Qiu-yao Dong[a,*], Hai-gang Wang[d], Yu Ren[a], Ze-yan Li[a], Yue-ying Du[c]

[a] *The Institute of Hydrogeology and Environmental Geology, Chinese Academy of Geological Science, China Geological Survey, Ministry of Natural Resources, Shijiazhuang* 050061, *China*
[b] *Key Laboratory of Groundwater Sciences and Engineering, Ministry of Natural Resources, Shijiazhuang*, 050061, *P.R. China*
[c] *North China University of Water Resources and Electric Power, Zhengzhou* 450011, *China*
[d] *China Institute of Geo-Environment Monitoring, Beijing* 100081, *China*

ARTICLE INFO

Article history:
Received 17 December 2022
Received in revised form 28 February 2023
Accepted 8 March 2023
Available online 22 March 2023

Keywords:
Landslide susceptibility model
Risk assessment
Machine learning
Support vector machines
Logistic regression
Random forest
Extreme gradient boosting
Linear discriminant analysis
Ensemble modeling
Factor analysis
Geological disaster survey engineering
Middle mountain area
Yellow River Basin

ABSTRACT

Landslide is a serious natural disaster next only to earthquake and flood, which will cause a great threat to people's lives and property safety. The traditional research of landslide disaster based on experience-driven or statistical model and its assessment results are subjective , difficult to quantify, and no pertinence. As a new research method for landslide susceptibility assessment, machine learning can greatly improve the landslide susceptibility model's accuracy by constructing statistical models. Taking Western Henan for example, the study selected 16 landslide influencing factors such as topography, geological environment, hydrological conditions, and human activities, and 11 landslide factors with the most significant influence on the landslide were selected by the recursive feature elimination (RFE) method. Five machine learning methods [Support Vector Machines (SVM), Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Linear Discriminant Analysis (LDA)] were used to construct the spatial distribution model of landslide susceptibility. The models were evaluated by the receiver operating characteristic curve and statistical index. After analysis and comparison, the XGBoost model (AUC 0.8759) performed the best and was suitable for dealing with regression problems. The model had a high adaptability to landslide data. According to the landslide susceptibility map of the five models, the overall distribution can be observed. The extremely high and high susceptibility areas are distributed in the Funiu Mountain range in the southwest, the Xiaoshan Mountain range in the west, and the Yellow River Basin in the north. These areas have large terrain fluctuations, complicated geological structural environments and frequent human engineering activities. The extremely high and highly prone areas were 12043.3 km^2 and 3087.45 km^2, accounting for 47.61% and 12.20% of the total area of the study area, respectively. Our study reflects the distribution of landslide susceptibility in western Henan Province, which provides a scientific basis for regional disaster warning, prediction, and resource protection. The study has important practical significance for subsequent landslide disaster management.

1. Introduction

Landslide is a complex geological evolution process induced by geologic structure, precipitation, and other internal and external factors. It is found at the junction of geological tectonic units, the region through large fracture zones, and the lowland of road excavation and slope collapse (Causes L, 2001). As the most serious geological hazard, landslide hazard has posed a severe threat to the safety of people's lives and property because of its wide distribution, sudden occurrence, colossal destruction and unpredictability (Dai FC et al., 2002; Yu FD et al., 2023). China is one of the countries suffering most seriously from landslide disaster in the world. In 2021, there were 4722 geological hazards in China, of which 2335 were landslide hazards, accounting for 49.4% of the total (Ministry of Natural Resources of the People 's Republic of China, 2022). Many methods for landslide activity monitoring have been used to reduce the enormous losses caused by landslides, and these methods have achieved specific effects. Early landslide hazard research mainly

First author: *E-mail address*: caowengeng@mail.cgs.gov.cn (Wen-geng Cao).

* Corresponding author: *E-mail address*: dongqiuyao@mail.cgs.gov.cn (Qiu-yao Dong).

Literary editor: Xi-jie Chen
doi:10.31035/cg2023013
2096-5192/© 2023 China Geology Editorial Office.

focused on landslide prediction and risk assessment through landslide hazard evaluation and landslide mapping (Guzzetti F, 2006). With the development of landslide monitoring technology, people use different methods, such as geodetic surveying, digital photogrammetry, remote sensing, etc. to monitor the structural deformation and surface displacement during landslides, and use the statistical analysis methods to predict landslide (Savvaidis PD, 2003). Traditional landslide hazard research is helpful in mastering and analyzing the characteristics of the landslide. However, it can only be used to monitor the proved landslide sites in the local scope, and has the defects of low monitoring efficiency and susceptibility to external conditions, and is unable to accurately predict the overall change characteristics of landslide at regional scale. With the development of digital technology and computer science, the scale of landslide research has changed from large-scale to fine-scale. In the last decade, the research direction has changed from monitoring the evolution of potential landslide to predicting landslide susceptibility. The research method has also changed from the analysis and statistics based on landslide data to the assessment of landslide susceptibility based on machine learning (Bao H et al., 2022).

Landslide susceptibility assessment is the basis of regional landslide risk assessment, prevention and control. It is essential to predict the spatial probability of landslide occurrence. Accurate landslide susceptibility assessment can provide adequate technical for disaster prevention and mitigation (Brabb EE, 1987). Based on the investigation of landslide hazards in the study area, researchers can obtain the landslide susceptibility assessment results by analyzing the internal and external factors affecting landslides, and statistical analysis of the landslide probability caused by the multiple landslide influencing factors under certain conditions. Based on the division of assessment units and the selection of environmental factors in the study area, an appropriate model is selected to evaluate the landslide susceptibility. Western Henan is in the transitional zone from the second step of the terrain to the third step. The geomorphic conditions are very complex, and it is located in the boundary zone between subtropical zone and warm temperate zone, and has significant differences in climate, vegetation, hydrology, and soil. With the continuous expand of urbanization, various large-scale engineering activities also increases the frequency of landslide hazards, which will directly threaten the safety of engineering activities, the ecological environment and people's lives and property. The intensity of landslide hazards in Henan Province is above the middle level in the whole country, and Western Henan is more susceptible to landslide in this province. For example, a large bedrock landslide in the east of the Xiaolangdi Reservoir Dam seriously affected the operation of water conservancy hub and traffic (Xu W et al., 2014). And in July 2007, heavy precipitation in Lushi County caused a landslide in north of Western Henan Grand Canyon, covered roads and trapping thousands of people inside the canyon.

Landslide sensitivity assessment based on the machine learning model can more accurately calculate the multivariate complex nonlinear relationship between landslide susceptibility and environmental factors. At the same time, it does not require the normal distribution of environmental factors and is suitable for large scale (Brenning A, 2005). Due to the uniqueness of landslide data in different regions, the applicability of models is different, so there is no universal applicability of optimal assessment model. In this study, various machine learning models were established after screening landslide susceptibility assessment factors, including Support Vector Machines (SVM), Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGBoost) and Linear Discriminant Analysis (LDA). Through establishing a landslide susceptibility model, this study simulated the spatial location of the potential landslide susceptibility area, and the model accuracy was verified by receiver operating characteristic (ROC) curve and statistical index to obtain the most suitable model for the data in the study area. The assessment of landslide susceptibility based on a machine learning model in Western Henan has important practical significance for regional landslide disaster management, which provides technical support for disaster warning and resource protection in Henan.

2. Geological Background

2.1. Geographical Environment

Western Henan refers to Luoyang and Sanmenxia region in the west of Henan Province, with a longitude of 110°21'18"–112 °58' 48" E and latitude of 33°34 '12 "–35 °4' 52" N, covering an area of 25539 km^2. It is in the middle latitude region and belongs to the boundary zone between subtropical and warm temperate zones. The annual average temperature in Western Henan is 13.8°C, gradually decreasing from south to north. It is a wet-subhumid monsoon climate, and the climate boundary is in the Funiu Mountain ridge within the region. The spatial variation of precipitation in Henan Province is very great, the precipitation is high in the south, less in the north, high in hilly and mountainous areas, and less in the plain. The annual precipitation in the mountainous areas of Western Henan is 700–900 mm. The Funiu Mountain area in Western Henan is the watershed of the Yangtze, Yellow and Huai Rivers, and the origin of most rivers in Henan Province. The rivers in the area are radially distributed from west to east, while the rivers cut deep into the terrain, creating favorable conditions for the landslide disasters.

2.2. Geomorphology

As an important geoenvironmental condition, the type and spatial distribution of geological geohazards are greatly controlled by the geomorphological factors, spatial distribution and their inter-combination. Western Henan is in the transition zone of the second and third steps, with

immense relief, complex geological environment and significant spatial-temporal differences in climate , which are prone to landslide hazards. The number of landslide hazards in the mountains and hills of Western Henan accounts for about 40% of the total landslide hazards in Henan Province. The main peaks in Western Henan are more than 1500 m above sea level, and some of them are more than 2000 m above sea level.

2.3. Regional Geology

The strata is fully developed in Western Henan, except for the Upper Ordovician, Upper Silurian and Lower Devonian, which are missing, and are exposed from the Archaeozoic to the Cenozoic. The structure of rock mass is fragile and prone to landslide hazards. The tectonic traces of the fractured zones affecting the western part of Henan are crisscrossed in a variety of ways. The neotectonic in Henan Province is characterized by vertical movements, which not only deform the strata but also evolve the paleogeography by the upward and downward movements of the crust. The surface formed by landslide hazards provides convenient conditions for developing subsequent disasters.

2.4. Hydrogeology

The major rivers in Henan all originate in the mountainous region of Western Henan, and the mountains and valleys are distributed alternately. In the terms of hydrogeology, the precipitation is the main factor in landslide hazards. The water-bearing rock groups in Western Henan mainly include loose rocks, carbonates, clastic rocks and fractured rocks such as metamorphic rocks and magmatic rocks. The distribution and migration of groundwater in fractured rocks such as carbonate rocks, metamorphic rocks and magmatic rocks distributed in Western Henan greatly influence the stability of rock and soil mass.

3. Research Methods

This study analyzed landslide susceptibility through six steps (Fig. 2): (1) Prepare the geospatial database. (2) Use the feature selection method to select appropriate landslide-affecting factors for landslide analysis. (3) Prepare training and testing datasets. (4) Construct landslide models. (5) Verify and comparing landslide models. (6) Draw landslide susceptibility maps (LSM).

3.1. Preparation of geospatial database

The study constructs a landslide spatial database as the data basis for analyzing the landslide hazards environment in Western Henan. The spatial database contains the spatial data of geological environment, landslide disaster points and landslide-influencing factors in the study area.

The landslide disaster points are extracted from the distribution data of geological disaster points in Western Henan through a geographic remote sensing ecological network. The landslide data was processed using the World Imagery Wayback tool in ArcGIS software to verify the accuracy of the landslide location based on historical map image data and landslide list. A total of 256 landslide points with a pixel size of 30 m×30 m were selected for landslide modeling analysis (Fig. 1). Landslide development is influenced by the combination of internal and external factors. The internal factors include rock mass structure and geotechnical properties, which play a controlling role in the slope ’s stability. The external factors include precipitation, rock weathering, human activities, etc. The terrain feature is one of the main influencing factors of landslide formation (Zheng X et al., 2021). According to the geological characteristics and historical landslide data in Western Henan, this study selected 16 landslide-affecting factors (slope angle, slope aspect, elevation, curvature, plan curvature, profile curvature, soil type, land cover, annual precipitation, lithology, distance to faults, distance to roads, distance to rivers, lineament density, road density and river density) based on the mechanism of the landslide (Table 1), and the typical influencing factors are shown in Fig. 3. The elevation, curvature, plan curvature, profile curvature, slope angle and slope aspect are extracted by 30m precision DEM digital elevation data using ArcGIS.

Based on landslide susceptibility analysis, different affecting factors were classified as follows (Fig. 3).

3.2. Model construction using machine learning algorithms

3.2.1. Support Vector Machines (SVM)

Support Vector Machine is a statistical learning method based on the principle of structural risk minimization and aiming at constructing the optimal hyperplane (Vapnik V, 1999). It can effectively process nonlinear data to improve classification observation results and is especially suitable for data processing with small sample sets (Noble WS, 2006).

The basic principles of SVM is that, for linear unfractional data $\{x_i, y_i\}; x_i \in R^d, y_i \in \{-1, +1\}$, i is the number of samples; d is the data dimension, and the original data needs to be mapped into a feature space by a nonlinear mapping φ(x). $\omega \cdot \varphi(x)+b=0$ is a hyperplane equation.In this case, the classification interval is equal to $2/\|\omega\|$. To maximize $2/\|\omega\|$, we can set $\|\omega\|^2/2$ minimized, and the classification line must satisfy the constraints:

$$y_i(\omega \cdot x_i + b) \geqslant 1 - \epsilon_i (\epsilon_i \geqslant 0)$$

Where: ϵ_i is the slack variable. While solving the classification hyperplane, the value ϵ_i is as small as possible (Oommen T et al., 2008).

3.2.2. Logistic Regression (LR)

The Logistic Regression model is used to handle an independent variable with multiple unrelated models of multivariate regression relationship between independent variables.[i] The independent variables in the logistic

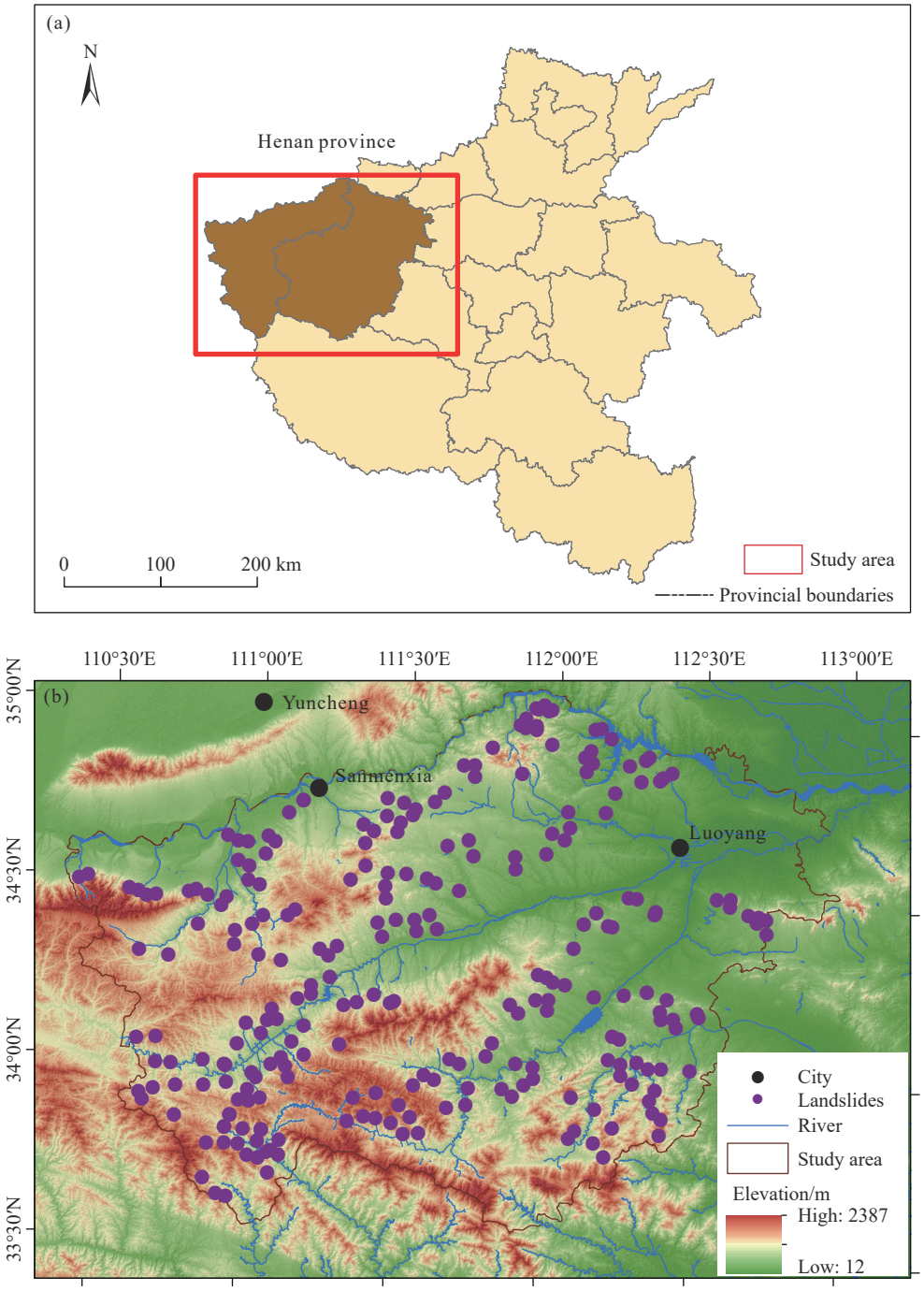

Fig. 1. Geographical location of the study area in Henan Province (a) and location of landslide in the study area (b).

regression model do not need to meet the normal distribution, and the independent variables are the assessment factors that affect the landslide unit and the non-landslide unit respectively. At the same time, the dichotomous problem of whether a landslide occurs is solved (Bui DT et al., 2011), which is between 0 and 1 (0 is the non-landslide unit, and 1 is the landslide unit). The probability of landslide occurrence is set as P, then the probability of landslide non-occurrence is Q= 1−P, and Logit transformation is performed on P. Then the regression equation is obtained,

$$\mathrm{LogitP} = a_0 + a_1X_{1j} + a_2X_{2j} + \cdots + a_nX_{nj}$$

that is,

$$P = \frac{\exp(a_0 + a_1X_{1j} + a_2X_{2j} + \cdots + a_nX_{nj})}{1 + \exp(a_0 + a_1X_{1j} + a_2X_{2j} + \cdots + a_nX_{nj})}$$

where, landslide probability P is the dependent variable, influencing factor set $[X_{1j}, X_{2j}, \cdots, X_{nj}]$ is the independent variable, and Logit P is the objective function of landslide probability, which is expressed as a linear combination of independent variables of each factor. $a_1, a_2, \cdots, a_n$are logistic regression coefficients, and a_0 is a constant, representing the logarithm value of the ratio between the probability of landslide occurrence and non-occurrence under the condition

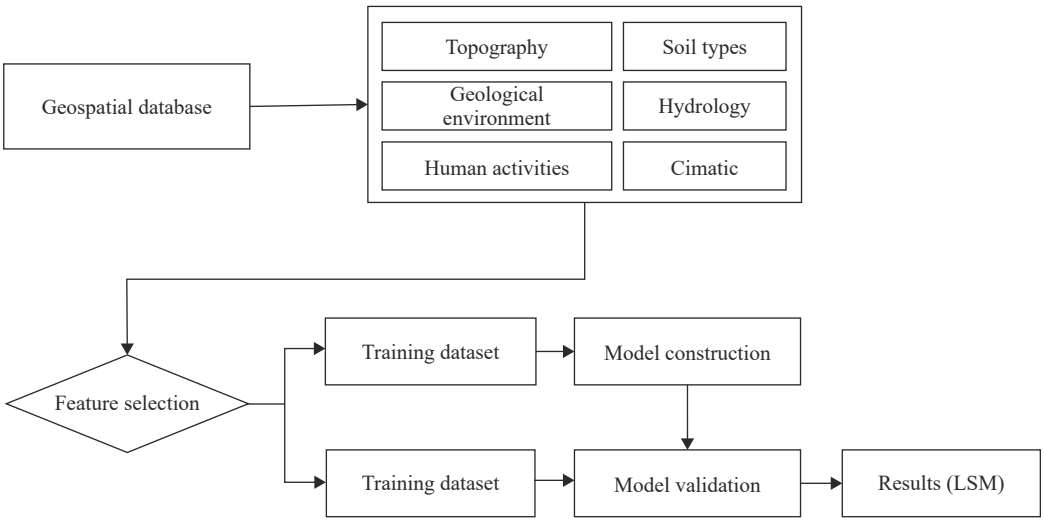

Fig. 2. Method adopted in the present study .

Table 1. Data sources of landslide points and landslide affecting factors.

Landslide affecting factors	Data acquisition	Data sources
Landslide points	Distribution data of geological disaster points in Western Henan	Geographic Remote sensing Ecological network (http://www.gisrs.cn)
Elevation	DEM digital elevation data with 30 m accuracy in Western Henan	Geospatial Data Cloud (http://www.gscloud.cn)
Slope		
Slope aspect		
Curvature		
Plan curvature		
Profile curvature		
Soil type	Soil type distribution data with 30 m precision in Western Henan	HWSD Soil Database
Land cover	Land cover data with 30 m accuracy in Western Henan	Global Soil Cover Database (http://www.globallandcover.com)
Annual precipitation	Precipitation data of 30 years in Western Henan	NCDC Exposes FTP Servers
Lithology	Lithologic data of Western Henan	Data Center for Resources and Environmental Sciences, Chinese Academy of Sciences (http://www.resdc.cn)
Distance to faults	Data of water system, traffic, settlement and land use with 30 m accuracy in Western Henan	National Geographic Information Resources Directory Service System (https://www.webmap.cn)
Distance to roads		
Distance to rivers		
Lineament density		
Road density		
River density		
Fault density		

that it is not affected by any landslide occurrence factor (Budimir MEA et al., 2015).

3.2.3. Random Forest (RF)

Based on the idea of parallel ensemble learning, Random Forest takes the decision tree as the basic model and constructs a set of decision tree models without strong dependence on each other by constructing different training datasets and feature spaces (Breiman L, 2001).

RF classification uses bootstrap sampling to extract k samples (generally 2/3) from the original training set T to generate a new training sample set and builds k decision tree models for each k samples to obtain k classification results. The classification error depends on the classification ability of each tree and the correlation between them. Finally, each record is voted on according to k classification results to determine its final classification. Many theoretical and empirical studies have proved that the random forest algorithms have high prediction accuracy, good tolerance to outliers and noise, and are one of the best machine learning classification and regression models (Tibshirani R, 1996).

3.2.4. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a machine-learning framework based on a gradient-lifting decision tree (Friedman JH, 2001). XGBoost expands and optimizes its structure, executes a quadratic Taylor expansion on the loss function, and uses the information of the first and second derivatives to automatically use the multi-thread parallel computation of the CPU during training (Chen T and Guestrin

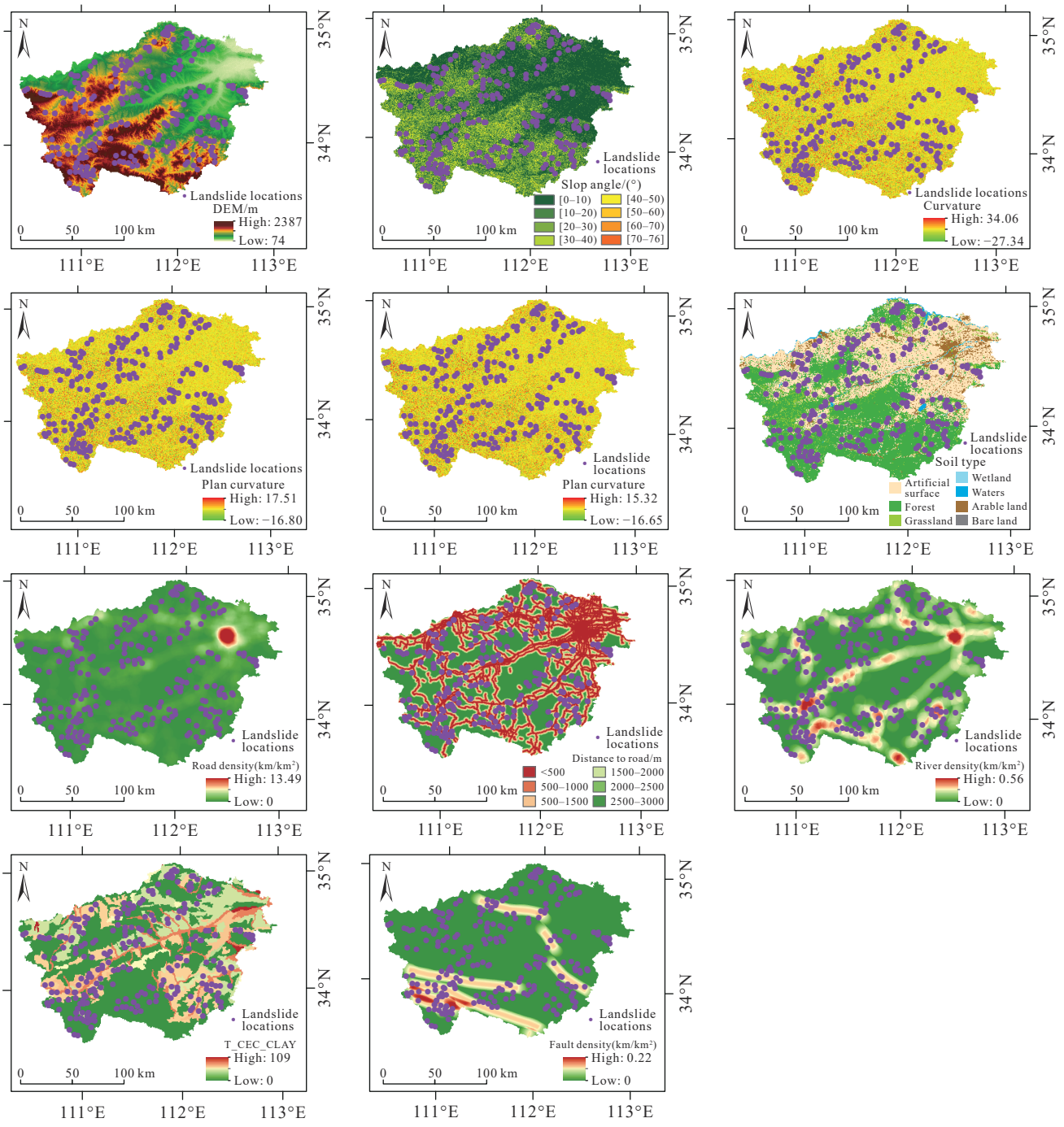

Fig. 3. Affecting factors of typical landslides.

G, 2016). In addition, to prevent over-fitting, XGBoost adds a regular penalty term to the loss function to reduce the complexity of the model and adopts the row and column sampling method to sample the model (Friedman JH, 2002). Its advantages are that computing resources are small, efficient and flexible, and easy to run (Fan Z et al., 2011).

The objective function is:

$$\pounds(\Phi) = \sum_i l(\bar{y}_i + y_i) + \sum_j \Omega(f_j)$$

$$\Omega(f_j) = \mathrm{Y}T + \frac{y}{2}w^2$$

where: Y is penalty coefficient, T is the number of leaf nodes in the CART tree, w is the weight of leaf nodes in each CART tree, l is the loss function, representing the error between the predicted value and the observed value, Ω is a regular penalty term function used to prevent overfitting, which can effectively limit the number of leaf nodes.

3.2.5. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis is a generalization of Fisher's linear discriminant method, which uses statistical, pattern recognition and machine learning methods to find a linear combination of features that can characterize two classes of objects or events or distinguish between them. The resulting combination can be used as a linear classifier to reduce the

dimensionality of subsequent classification (Sharma A and Paliwal KK, 2015). Based on a given set of training samples, linear discriminant analysis tries to project the samples onto a straight line so that the projection points of similar samples are as close as possible, and those of different samples are as far away. When classifying the new samples, they are projected onto the same line, and then the category of the new samples is determined according to the position of the projection points.

3.3. Assessment and comparison methods

The training and testing data sets were used to compare the spatial prediction ability of each landslide model. The training data set is used for modeling to reflect the degree of fitting of the model to the data, while the test data set is used to reflect the model's prediction ability. In this study, the performance of the five landslide models was compared by statistical index-based assessment and ROC curve.

3.3.1. Statistical index based assessments

In this study, the performance of the landslide models was verified by determining the statistical indicators: the sensitivity, specificity, accuracy and root mean square error. Sensitivity refers to how many positive examples in the sample are predicted correctly by the landslide models, which indicates the predictive ability of the landslide models for the classification of landslide pixels (Bennett ND et al., 2013). Specificity refers to how many negative examples are in the sample after the landslide models are predicted correctly, which indicates the predictive ability of the landslide models for non-landslide pixel classification. Accuracy is the proportion of correctly classified landslide and non-landslide pixels, indicating the performance of the landslide models. And root means square error shows the error measure of the same unit as the original data. The smaller the RMSE value, the better the performance of the landslide models.

3.3.2. Receiver operating characteristic (ROC) curve

ROC curve is a method to verify the landslide model based on the confusion matrix. It is a curve drawn according to several different binary classification limit worth (thresholds), with the rate of increase (sensitivity, TPR) as the ordinate and the false positive rate (1-specificity, FPR) as the abscissa. ROC curve can easily detect any threshold's influence on the learner's generalization performance, which helps select the best threshold. The closer the ROC curve is to the upper left, the higher the accuracy of the model, and the point on the ROC curve closest to the upper left is the best threshold with the fewest classification errors. The area enclosed by the ROC curve and the coordinate axes (AUC) can directly reflect the classification ability expressed by the ROC curve. The closer the AUC is to 1.0, the higher the authenticity of the detection method is. It has no authenticity and application value when it is equal to 0.5. The ROC curve is simple and intuitive, and the accuracy can be judged and analyzed by the naked eye (Cantarino I et al., 2019).

4. Model study and analysis

4.1. Selecting landslide affecting factors

In this study, a total of 16 geological and environmental factors (slope, aspect, elevation, curvature, plane curvature, profile curvature, soil type, land cover, annual precipitation, lithology, distance to linear structure, distance to road, distance to river, line density, road density and river density) were selected as landslide influencing parameters. However, these geological and environmental factors may have different influence on the landslide model. Now, the recursive feature elimination (RFE)method is used to repeatedly build the model to evaluate these landslide influencing parameters, eliminate unimportant or irrelevant factors, further fit the critical influencing factors, and finally select the best feature subset in the classification. The RFE method needs to build a training classifier, calculate the importance measure of features, remove the irrelevant features with low importance measure, and then repeat this process until the best feature subset is selected (Munasinghe K and Karunanayake P, 2021). The RFE method obtained the relative importance of each influence factor (Fig. 4). The higher the value, the stronger the impact on the landslide model. Finally, 11 landslide affecting factors (slope angle, elevation, curvature, plan curvature, profile curvature, land cover, lithology, distance to roads, distance to faults, road density and river density) were selected for landslide modeling.

4.2. Training and validating landslide models

ROC is used in machine learning to judge the merit of classification and detection results. It is mainly analyzed by ROC curves on a two-dimensional plane. The horizontal coordinate of the plane is the false positive rate (FPR) and the vertical coordinate is the true positive rate (TPR). It can be mapped to a point on the ROC plane based on the performance of the classifier on the test sample. By adjusting the threshold used in classifying this classifier, we get a curve that passes through (0, 0), (1, 1), which is the ROC curve of this classifier. The two most important metrics within the ROC are，

True positive rate： $\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$,

False positive rate： $\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}}$,

Where, TP denotes positive samples predicted by the model as positive class; FN denotes positive samples predicted by the model as negative class; FP denotes negative samples predicted by the model as positive class; TN denotes negative samples predicted by the model as negative class. TPR focuses on the correct rate and FPR focuses on the error rate. Ideally, the larger the TP and TN, the better, and the smaller the FP and FN, the better. That is, in the ROC graph, the closer the point to (0, 1) corresponds to the better classification performance of the model.

The proposed prediction model for landslide susceptibility

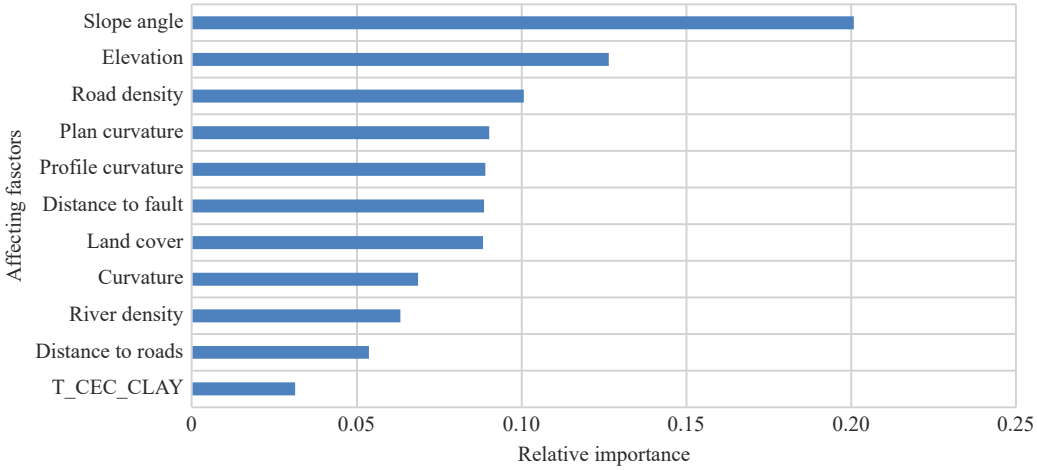

Fig. 4. Importance ranking of landslide affecting factors based on RFE.

was constructed by generating training and testing datasets. The training dataset is used to train the landslide models, while the test dataset is used to verify the performance of the landslide models. The 256 landslide points were randomly assigned into two parts. 70% of the points were used as the training dataset, and the remaining 30% were used as the test dataset. Then 70% of the non-landslide data set was selected to generate the training dataset and 30% to construct the test dataset. Landslide influencing factor data shall be sampled to generate the final data set. The training and testing datasets were used to analyze the performance of five machine learning methods. The results showed that the ROC curves of different models were analyzed using the training datasets. The highest values belong to the RF model and XGBoost model (AUC 1.0000), followed by the SVM model (AUC 0.9844), LDA model (AUC 0.8903) and LR model (AUC 0.8796). Using the test dataset to analyze the ROC curves of different models (Fig. 5), it can be observed that the highest value belongs to the XGBoost model (AUC 0.8759), then RF model (AUC 0.8743), LDA model (AUC 0.8685), LR model (AUC 0.8624) and SVM model (AUC 0.7666) were followed. It can be analyzed that the XGBoost model has the best predictive ability.

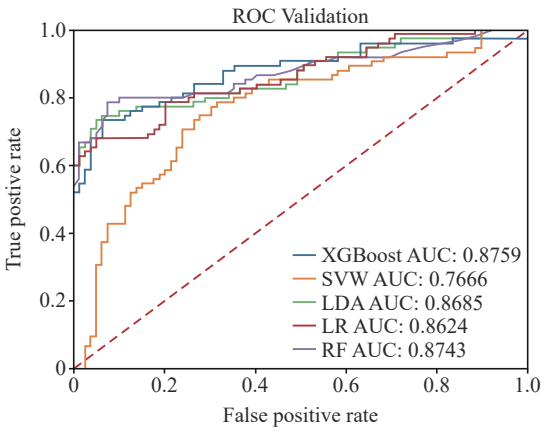

Fig. 5. The ROC curves of different landslide models using testing dataset.

4.3. Drawing landslide susceptibility maps

The landslide susceptibility were mapped based on five landslide models. By generating a landslide susceptibility index (LSIs), a unique sensitivity index is assigned to each pixel in the study area. Then the geometric interval (GI) tool of ArcGIS was used to reclassify LSIs into different intervals. Based on the LSIs interval, five susceptibility grades were determined: Very low, low, moderate, high and very high for dividing landslide susceptibility (Fig. 6).

In addition, five landslide sensitivity maps were combined with slope maps to determine slope intervals more prone to landslides (Fig. 7). The overlay analysis of results confirms that most landslides have occurred in the susceptibility grades of very high and high , which are associated with the moderate slope of the ground varying from 10° to 30°.

5. Discussions

Compared with traditional landslide susceptibility prediction, machine learning methods have better nonlinear prediction ability in solving many practical problems. This study evaluated and compared five machine learning methods (SVM, LR, RF, XGBoost and LDA). Among them, RF, SVM, and LR have been widely used in the spatial prediction of landslide susceptibility. XGBoost and LDA methods have also been successively applied in geological hazard assessment. In this study, recursive feature elimination method (RFE) was used to evaluate the impact degree of landslide influencing parameters, eliminate unimportant factors, further fit important influencing factors, and finally select the best feature subset in the classification. The results showed that among the 16 landslide impact factor, 10 factors (slope angle, elevation, curvature, plan curvature, profile curvature, land cover, lithology, distance to roads, road density and river density) are relatively high importance to the landslide prediction model, so these factors are used to train the landslide model. The accuracy was verified by the ROC curve and statistical index-based method. And after analysis and comparison, it is concluded that the XGBoost model

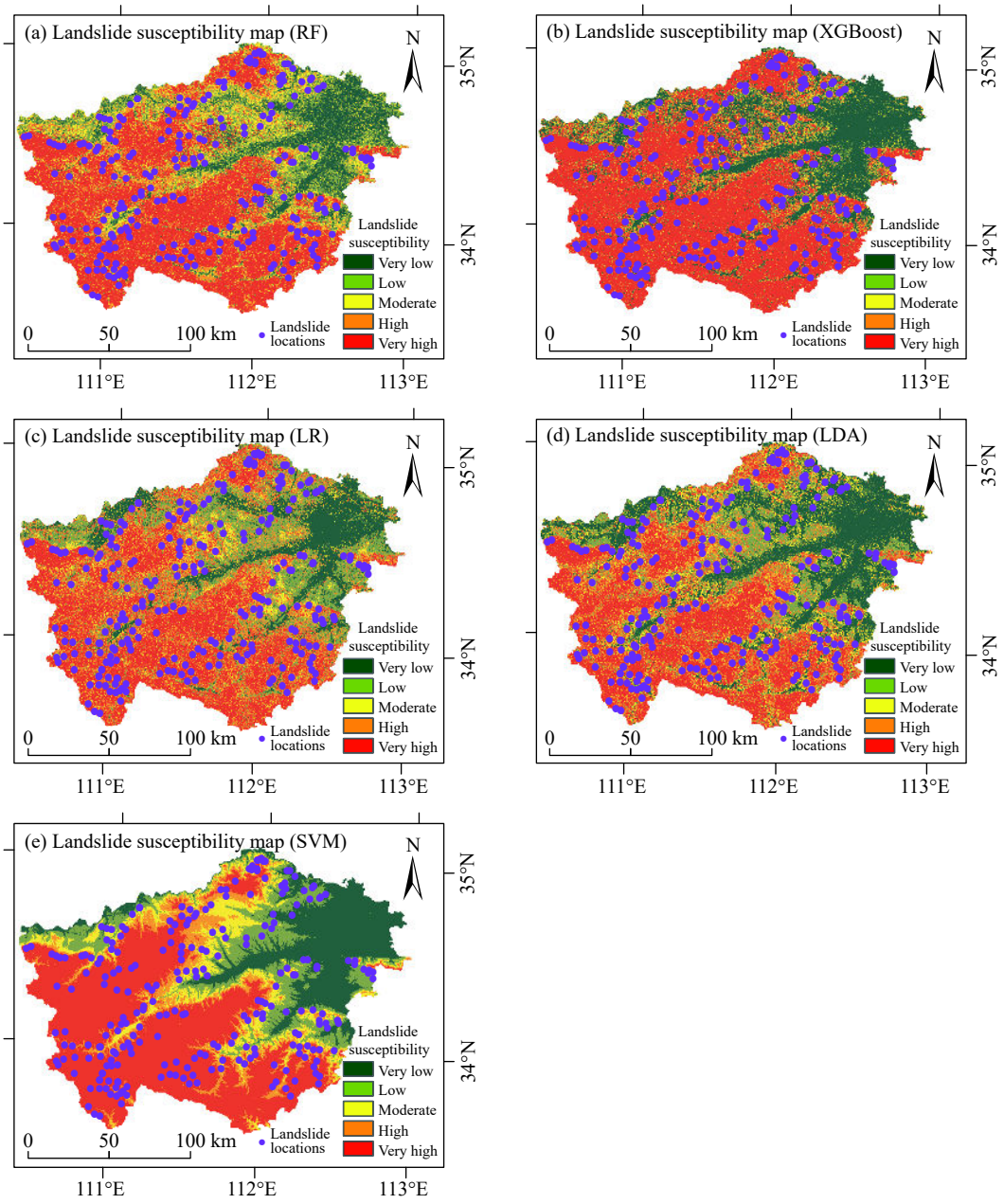

Fig. 6. Landslide susceptibility maps of different landslide models.

(AUC 0.8759) has the best performance and is suitable for dealing with regression problems and has high adaptability to landslide data. However, the XGBoost model needed to consider the situation of time consumption and insufficient memory when training big data. RF model (AUC 0.8743) also performs well. When the amount of data is large, and the processing speed is considered, the RF model can be used to solve practical problems.

6. Conclusions

According to the landslide susceptibility map of the five landslide models, the overall distribution trend can be observed. The extremely high and high susceptibility areas are distributed in the Funiu Mountain range in the southwest, the Xiaoshan Mountain range in the west and the Yellow River basin in the north. These areas have large terrain fluctuations, complicated geological structural environments and frequent human engineering activities. The moderately-prone areas are distributed in valleys such as the Luohe and Yinhe rivers. The low- and very-low-prone areas are distributed in the eastern part of Luoyang Basin, where the terrain is flat and open, and landslide hazards are rare. The prediction results of the XGBoost model and RF model were compared and analyzed. According to the XGBoost prediction distribution map, landslide development was concentrated in high–extremely high prone areas, the spatial distribution of landslide-prone areas was tree-like, and high–extremely high prone areas were concentrated in mountainous and hilly areas. According to the distribution map of RF prediction, the landslide development is scattered in the high–extremely high prone area. In the western Xiaoshan Mountains, where landslides are easily

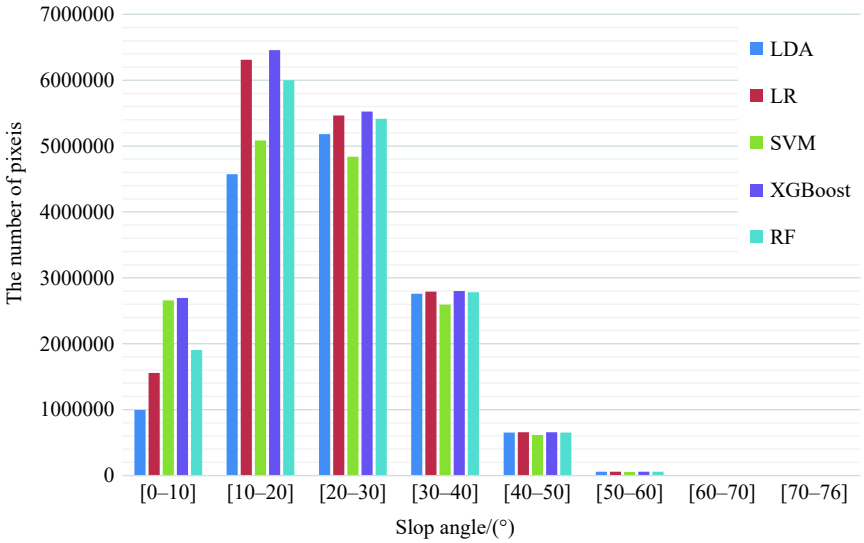

Fig. 7. Distribution of pixels of high and very high classes on slope map.

induced, the region ’s edge areas are divided into medium-low-prone areas. Its accuracy is not as good as that of the XGBoost model, so the XGBoost model is more suitable for the landslide susceptibility assessment in this region. Therefore, in the practical application of landslide susceptibility spatial prediction, the XGBoost model can be used to evaluate and develop a better landslide susceptibility map. On this basis, appropriate landslide disaster management can be carried out.

After years of geological hazard risk assessment development, traditional landslide hazard research is often based on empirical driving or theoretical statistical models, and the assessment results are subjective and difficult to quantify. Machine learning can combine landslide hazards and affecting factors through data processing, solve the nonlinear relationship in the affecting factors, and greatly improve the precision and accuracy of the landslide susceptibility model. Based on a variety of machine learning model, this study has better predicted landslides liability distribution of Western Henan, and analyzed that the XGBoost model is most suitable for the landslide liability assessment in Western Henan, which provides a scientific basis for disaster warning and prediction and resource protection in Henan Province, and has important practical significance for regional landslide disaster management.

CRediT authorship contribution statement

Cao-wen Geng, Fu Yu and Dong-qiu Yao conceived of the presented idea. Cao-wen Geng and Dong-qiu Yao 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 work was financially supported by National Natural Science Foundation of China (41972262) ,Hebei Natural Science Foundation for Excellent Young Scholars (D2020504032), Central Plains Science and technology innovation leader Project (214200510030) and Key research and development Project of Henan province (221111321500).

References

Bao H, Zeng CY, Peng Y, Wu SH. 2022. The use of digital technologies for landslide disaster risk research and disaster risk management: progress and prospects. Environmental Earth Sciences, 81(18), 446–456. doi: 10.1007/s12665-022-10575-7.

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

Brabb EE. 1987. Innovative approaches to landslide hazard and risk mapping. 307–324. doi: 10.1016/0148-9062(87)91363-5.

Breiman L. 2001. Random forests. Machine Learning, 45, 5–32. doi: 10.1023/A:1010933404324.

Brenning A. 2005. Spatial prediction models for landslide hazards: review, comparison and evaluation. Natural Hazards and Earth System Sciences, 5(6), 853–862. doi: 10.5194/nhess-5-853-2005.

Bui DT, Lofman O, Revhaug I, Dick O. 2011. Landslide susceptibility analysis in the Hoa Binh province of Vietnam using statistical index and logistic regression. Natural Hazards, 59, 1413–1444. doi: 10.1007/s11069-011-9844-2.

Budimir MEA, Atkinson PM, Lewis HG. 2015. A systematic review of landslide probability mapping using logistic regression. Landslides, 12, 419–436. doi: 10.1007/s10346-014-0550-5.

Cantarino I, Carrion MA, Goerlich F, Ibañez VM. 2019. A ROC analysis-based classification method for landslide susceptibility maps. Landslides, 16, 265–282. doi: 10.1007/s10346-018-1063-4.

Causes L. 2001. Landslide types and processes. US Geological Survey: Reston. VA. USA. 10.

Chen T, Guestrin C. 2016. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785–794. doi: 10.1145/2939672.2939785.

Dai FC, Lee CF, Ngai YY. 2002. Landslide risk assessment and

management: an overview. Engineering Geology, 64(1), 65–87. doi: 10.1016/s0013-7952(01)00093-x.

Fan Z, Xu Y, Zhang D. 2011. Local linear discriminant analysis framework using sample neighbors. IEEE Transactions on Neural Networks, 22(7), 1119–1132. doi: 10.1109/tnn.2011.2152852.

Friedman JH. 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189–1232. doi: 10.1214/aos/1013203451.

Friedman JH. 2002. Stochastic gradient boosting. Computational statistics & data analysis, 38(4), 367–378. doi: 10.1016/s0167-9473(01)00065-2.

Guzzetti F. 2006. Landslide Hazard and Risk Assessment. Transportation Research Board Special Report. 373.

Ministry of Natural Resources of the people's Republic of China. 2022. National Geological Disaster Situation in 2021 and Geological disaster Trend forec ast in 2022. mnr.gov.cn/dt/ywbb/202201/t20220113_2717375.html.

Munasinghe K, Karunanayake P. 2021. Recursive feature elimination for machine learning-based landslide prediction models. 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), IEEE, 126–129. doi: 10.1109/icaiic51459.2021.9415232.

Noble WS. 2006. What is a support vector machine. Nature biotechnology, 24(12), 1565–1567. doi: 10.1038/nbt1206-1565.

Oommen T, Misra D, Twarakavi NKC, Prakash A, Sahoo B, Bandopadhyay S. 2008. An objective analysis of support vector machine based classification for remote sensing. Mathematical geosciences, 40, 409–424. doi: 10.1007/s11004-008-9156-6.

Savvaidis PD. 2003. Existing landslide monitoring systems and techniques. In Proceedings of the conference from stars to earth and culture, The Aristotle University of Thessaloniki. Thessaloniki, Greece, 242-258.

Sharma A, Paliwal KK. 2015. Linear discriminant analysis for the small sample size problem: an overview. International Journal of Machine Learning and Cybernetics, 6, 443–454. doi: 10.1007/s13042-013-0226-9.

Tibshirani R. 1996. Bias, variance and prediction error for classification rules. University of Toronto, Department of Statistics. 13.

Vapnik V. 1999. The nature of statistical learning theory. Springer Science & Business Media. 314. doi: 10.1007/978-1-4757-3264-1.

Xu WJ, Jie YX, Li QB, Wang XB, Yu YZ. 2014. Genesis, mechanism, and stability of the Dongmiaojia landslide, yellow river, China. International journal of rock mechanics and mining sciences, 67, 57–68. doi: 10.1016/j.ijrmms.2014.01.010.

Yu FD, Qiao G, Wang K, Zhang X. 2023. Investigation of groundwater characteristics and its influence on Landslides in Heifangtai Plateau using comprehensive geophysical methods. Journal of Groundwater Science and Engineering, 11(2), 171–182.

Zheng XX, He GJ, Wang SS, Wang Y, Wang GZ, Yang ZY, Yu JC, Wang N. 2021. Comparison of machine learning methods for potential active landslide hazards identification with multi-source data. ISPRS International Journal of Geo-Information, 10(4), 253–274. doi: 10.3390/ijgi10040253.