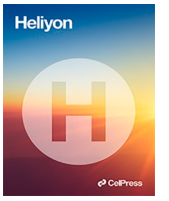
Heliyon 9 (2023) e21542

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



A python system for regional landslide susceptibility assessment by integrating machine learning models and its application

Zizheng Guo [a,b,c], Fei Guo [a,b,*], Yu Zhang [d,e], Jun He [c,**], Guangming Li [f], Yufei Yang [c], Xiaobo Zhang [g]

[a] *Hubei Key Laboratory of Disaster Prevention and Mitigation (China Three Gorges University), Yichang, 443002, China*
[b] *Key Laboratory of Geological Hazards on Three Gorges Reservoir Area (China Three Gorges University), Ministry of Education, Yichang, 443002, China*
[c] *School of Civil and Transportation Engineering, Hebei University of Technology, Tianjin, 300401, China*
[d] *Zhejiang Geology and Mineral Technology Co. LTD, Hangzhou, 310007, China*
[e] *Wenzhou Engineering Survey Institute Co., LTD, Wenzhou, 325006, China*
[f] *Tianjin Municipal Engineering Design & Research Institute (TMEDI), Tianjin, 300392, China*
[g] *Beijing Glory PKPM Technology Co.,Ltd., Beijing, 100013, China*

ARTICLE INFO

Keywords:
Landslide susceptibility assessment
Python
Machine learning models
Loess plateau
GIS

ABSTRACT

Landslide susceptibility assessment is considered the first step in landslide risk assessment, but current studies mostly rely on GIS platforms or other software for data preprocessing. The modeling process is relatively complicated and multi-models cannot be integrated. With regard to this issue, this study develops a Python system for automatic assessment of regional landslide susceptibility. The Python system implements landslide susceptibility assessment through three modules: geographic data processing, machine learning modeling and result evaluation analysis. For geographic data processing, ten landslide influencing factors can be used to construct an evaluation factor dataset and reclassify the thematic maps based on the frequency ratio method. Four built-in machine learning models (logistic regression (LR), multi-layer perceptron (MLP), support vector machine (SVM) and extreme gradient boosting (XGBoost)) are integrated into the system to complete susceptibility modeling and calculation. Additionally, receiver operating characteristic (ROC) curves can be automatically generated to evaluate the accuracy. The system was then applied into Lantian County in Shaanxi Province as a demonstration example. The results show that the areas under the ROC curve (AUC) of the four models are 0.838 (LR)、0.882 (SVM)、0.809 (MLP) and 0.812 (XGBoost), respectively, indicating that the SVM model was the most suitable model for landslide susceptibility assessment in Lantian County in the Loess Plateau of China. The system has now been made open source on Github, which can effectively improve the efficiency of regional landslide susceptibility assessment, especially provide tools for data processing and modeling for non-professionals.

* Corresponding author. Hubei Key Laboratory of Disaster Prevention and Mitigation (China Three Gorges University), Yichang, 443002, China.
** Corresponding author.
E-mail addresses: ybbnui.2008@163.com (F. Guo), 202111601003@stu.hebut.edu.cn (J. He).

https://doi.org/10.1016/j.heliyon.2023.e21542
Received 4 July 2023; Received in revised form 23 October 2023; Accepted 23 October 2023
Available online 2 November 2023

1. Introduction

Landslides are the most frequent and costly geological disaster worldwide which cause considerable casualties and property losses on a yearly basis [1–4]. Drawing a landslide susceptibility map can reveal where landslides are more likely to occur in the future through predict the spatial probability of landslides, thus providing further basis and reference for landslide risk assessment [5,6]. Conducting fast and efficient automated landslide susceptibility assessment tools has therefore been an essential tool for disaster prevention and mitigation for local authorities and stakeholders.

Commonly used approaches for landslide susceptibility assessment have been developed in recent years, which can be roughly included into groups: expert empirical models, physically-based models, statistical models and machine learning (ML) models [7–11]. Multiple comparative studies [12,13] have shown that statistical and ML models typically have higher precision and accuracy and are therefore more widely used for the task of landslide susceptibility assessment. Common statistical models include principal component analysis (PCA), informativeness, and weight of evidence [5,14]. With the development of computer science, such studies now prefer to use the latest machine learning (ML) methods with more complex structures to assess the landslide susceptibility at a regional scale, such as artificial neural networks (ANN), random forests (RF), support vector machines (SVM), K-Nearest Neighbors (KNN) and Naive Bayes (NB) [15–18]. Table 1 gives the machine learning techniques commonly used for landslide susceptibility assessment, and compare their advantages and shortcomings. In addition, some recent studies have developed integrated (hybrid) ML models for landslide susceptibility assessment, which have also achieved satisfactory results: for example, Chen and Li [19] composed different hybrid ML models by combining error reduction pruning tree models with multiple sampling techniques (Bagging, Dagging, Real Adaboost) and found that the hybrid models had higher accuracy than the single ML model. Compared to statistical methods which emphasize explanation and inference to explore causal relationships behind observed phenomena, ML-based models focus on discovering complex decision rules and patterns from data for generalization and prediction. Thus, novel relationships between the correlate factors and landslide susceptibility can be discovered through these ML techniques.

Landslide susceptibility assessment at a regional scale relies on establishing strong links between the occurrence of landslides the topography, landform and external environment of its location. The research objectives of data-driven models for regional landslide susceptibility assessment usually can be included into the following three parts: (1) statistical pattern analysis of influencing factors and historical landslides, sampling and data set construction; (2) landslide susceptibility modeling, result analysis and validation; and (3) landslide susceptibility mapping [20,21]. These steps generally involve multiple software or tools. Among of them the Geological information system (GIS) provides an excellent platform for geographic information analysis and statistics [22]. Hence, GIS has been an essential and basic tool to conduct the statistical analysis for landslide influencing factors. In addition, interpreting these involved factors as accurately as possible to enable the calculation of susceptibility models requires the implementation through computational science software such as Matlab [23], Python [24] and SPSS Modeler [25]. Furthermore, the validation of modeling results needs to be implemented in cooperation with statistical software and GIS platforms. The intersection of these steps makes the process of landslide

Table 1
Summary of machine learning models commonly applied in landslide susceptibility assessment [16].

Methods	Advantages	Disadvantages
Linear Models (e.g. LR)	1. Easy to interpret and understand. 2. Suitability for binary and multiclass classification problems. 3. It is robust to outliers and noise in the data. 4.Does not require input features to be scaled. 5.Does not require any tuning.	1. Fails to solve complex non-linear relationships/problems. 2. It does not handle missing values well. 3.Requires data preprocessing. 4. It suffers from overfitting if the model is too complex or the number of features is too high.
SVM	1. Flexible and can handle non-linearly separable data. 2. It provides a clear separation boundary between classes. 3. It is robust to noise and irrelevant features.	1. High computational complexity, especially for large datasets. 2. Not suitable for very high dimensional data. 3. Performance is affected by hyperparameters.
Ensemble methods (e.g. XGBoost, RF)	1. Can reduce the variance of the predictions and thus make them more robust to errors. 2. Can handle higher dimensionality data. 3. Do not require any assumptions about the underlying distribution of the data. 4. These models are relatively easy to understand and interpret compared to more sophisticated machine learning models.	1. Due to the lack of constraints, Ensemble methods may not capture all the nuances in the data. 2. As the number of variables increases, the computational complexity of these models also increases linearly. 3. Sensitive to outliers that can significantly impact their performance.
Neighbor-based methods (e.g. KNN)	1. KNN is a non-parametric model that makes no assumptions about the distribution of the data. 2. The “distance” to neighbors can be interpreted as a measure of similarity, providing intuitive insights.	1. Hard to reproduce results elsewhere. 2. As the number of samples or features increases, the curse of dimensionality comes into play, slowing down both training and prediction times. 3. KNN can easily overfit to the training data if it has many outliers or if it uses a small training set.
Naive Bayes	1. Can be extremely fast compared to more sophisticated methods. 2. Bayesian models are robust to noisy data as they incorporate prior information about the parameters. 3. The structure of a Bayesian model can be easily modified to accommodate prior knowledge or new data.	1. Computing the posterior distribution for large datasets can be computationally intensive. 2. The choice of hyperparameters can significantly impact the performance of Bayesian models, which can be difficult and time-consuming.

susceptibility assessment, especially the pre-processing of data, complex and inefficient. A few scholars have tried to develop software or embed toolkits (plug-ins) in GIS to perform some of the tasks in the susceptibility assessment process. For example, Bartolini et al. [26] developed a QVAST plug-in that can be used to analyze volcanic susceptibility but was not applicable to landslide hazards; Alvioli et al. [27] built a toolkit that can be used to classify slope units and the results can be used as an intermediate step in landslide susceptibility assessment; Sahin et al. [28] and Torizin et al. [24] developed an open source toolkit for generating regional landslide susceptibility maps using R and Python, respectively, but the former only had two models for users to choose from: the logistic regression model (LR) and RF. Guo et al. [29] developed a plug-in called FSLAM for shallow landslides susceptibility assessment based on an open source GIS platform, but due to the using a physically based model, the accuracy of the results was lower than that of the commonly used data-driven models. As can be seen from the literature, previous research has been devoted to promoting more integrated and efficient tools for landslide susceptibility assessment, but most have focused on assessing the effectiveness of landslide susceptibility modeling methods and comparing the strengths and weaknesses between them. However, there is still a lack of automated tools that can be used in the whole process of landslide susceptibility assessment on the regional scale, and especially studies on the application of automated tools in specific tasks. An automated system that integrates advanced landslide susceptibility assessment techniques has been an important and urgent tool for stakeholders.

In response to the shortcoming of the current automatic landslide susceptibility assessment system, the purpose of this study is to develop a GIS software-independent landslide susceptibility assessment system based on Python language, which can automatically complete the regional landslide susceptibility assessment and its accuracy analysis through four different built-in models after the user inputs the influence factor files. The Lantian country in the Shanxi province of China, which suffers heavy threats from frequently occurring landslides, was selected as the study region.

2. Methods

2.1. Landslide susceptibility evaluation methods

The process of modeling landslide susceptibility involves two main aspects: statistical analysis of influencing factors and construction landslide susceptibility models. The former uses statistical methods to classify influencing factors into several sub-intervals and evaluates the impact of different subcategories on landslide occurrence. The latter introduces susceptibility evaluation models to learn the non-linear relationship between landslide influencing factors and historical landslides. Subsequently, the evaluation models are used to predict the probability of landslides occurring in the whole region. Specificity, for the first aspect, this study uses a frequency ratio (FR) method [30] to count the relationship between historical landslides and the influencing factors. As comes to the second aspect, although some studies have demonstrated the satisfactory accuracy of ML models for regional landslide susceptibility assessment, there is no consensus on the prediction performance and application scope for various ML models in the literature [31,32]. Hence, in order to increase the diversity of modeling methods for landslide susceptibility assessment, a total of four commonly used ML models are built into this system, namely LR, multilayer perceptron (MLP), SVM and extreme gradient boosting (XGBoost).

2.1.1. Frequency ratio analysis

The frequency ratio (FR) statistical method has been widely used in landslide susceptibility assessment, and its main function is to be able to divide continuous-type data into more reasonable subintervals [12,33]. For each particular landslide influencing factor, its landslide frequency ratio in each sub-interval is calculated as follows:

$$FR = \frac{N_i}{N_0} \Big/ \frac{S_i}{S_0} \tag{1}$$

where N_i denotes the number of landslide events in subclass i of an influencing factor, N_0 is the total number of landslides, S_i means the covered area of subclass i of this factor, S_0 is the total area of the study area. When the FR value of an interval is greater than 1, it means that the distribution density of landslides within the range is greater than the proportion of the interval in the whole area, i.e. the interval is favorable for landslides to occur; conversely, when the FR value of the interval is less than 1, it means that the density of landslides in the interval is low and it is an unfavorable area for landslides to occur.

2.1.2. Landslide susceptibility evaluation models

2.1.2.1. Logistic regression model (LR). The LR is a commonly used statistical method that is based on a linear regression model by transforming the real values of the latter's output into values between {0, 1} by means of a transformation function, as follows [5]:

$$z = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n \tag{2}$$

$$P = \frac{1}{1 + e^{-z}} \tag{3}$$

where P is the probability of a landslide occurring, x_i ($i = 1, 2, \ldots, n$) is the influencing factor, a_i ($i = 1, \ldots, n$) is the regression coefficient and a_0 is the intercept, n is the number of evaluation factors.

2.1.2.2. Multi-layer perceptron (MLP). The MLP has a similar structure to the ANN model, both consisting of an input layer, hidden layers and an output layer. The nodes in each layer are connected to the next layer, with a non-linear activation function for each hidden layer node [34]. The layers in models are linked to each other by weights and these weights were optimized using a back-propagation algorithm based on error gradients. The linear rule is changed through the multilayer activation functions, which is the reason why the MLP model is able to identify non-linear relationship between input data with targets. In addition, the hyper-parameter settings of the MLP model include the number of hidden layers, the number of nodes in each hidden layer, the activation function and the optimization algorithm of loss function. All of the hyperparameters can influence the final output values of the MLP. Therefore, in order to obtain a more efficient and robust performance, the system designed in this study also includes the hyper-parameters optimization and adjustment of the algorithm structure.

2.1.2.3. Support vector machine (SVM). The principle of the support vector machine algorithm is to map the original sample space into a higher dimensional linear space by means of a non-linear mapping. In the higher dimensional space, the SVM can create a classification hyperplane $W{\cdot}\Phi(x)+b=0$, where W is the plane normal vector and b is the intercept. So, it can achieve linear separability of the input samples in the higher dimensional space. The ultimate aim of the SVM is to find the maximized interval hyperplane for samples which can classify the landslides and non-landslides. Where the sample interval ($2/\|w\|$) is provided by several samples from different classes called support vectors. The corresponding solution to maximize ($2/\|\mathrm{w}\|$) can be transformed into a minimization problem as follows [35]:

$$min:\frac{1}{2}\|\ \mathrm{W}\|^2+c\sum_{i=1}^{m}\xi_i \tag{4}$$

$$s.t.y_i(W\cdot\Phi(x_i)+b)\geq 1-\xi_i,\xi_i\geq 0,i=0,1,2,\ldots,m \tag{5}$$

where c is the penalty factor, ξ_i is the relaxation factor, the classification error is introduced for linearly inseparable problems and m is the number of samples. To solve the above optimization problem, the SVM model introduces a Lagrangian function to transform this optimization problem into its dual form for solution, and eventually obtains its decision function for the optimal classification hyperplane, thus achieving classification.

2.1.2.4. Extreme gradient boosting (XGBoost)). XGBoost is a machine learning algorithm developed in recent years, which is an integrated model guided by the idea of Boosting integration, using Classification and Regression Trees (C&RT) as the basic classifier [36]. The algorithm adds a new C&RT tree per iteration to learn the residuals between the existing model and the objective values. Thus, XGBoost can improve the accuracy of the existing model step by step through multiple iterations. When n sub-predictors are trained, the scores of the leaf nodes are calculated based on the sample features and the predicted values of the samples are obtained by accumulating the results of the sub-predictors. Thus, the regularization objective function for the t-th iteration in the XGBoost model is:

$$obj^t=\sum\nolimits_{i=1}^{n}l\left(y_i,\widehat{y}_i^{(t-1)}+f_t(x_i)\right)+\Omega(f_t) \tag{6}$$

$$\Omega(f_t)=\gamma T+\frac{1}{2}\lambda\|W\|^2 \tag{7}$$

where $l(y_i,\widehat{y}_i^{(t-1)}+f_t(x_i))$ is the loss function, $\Omega(f_t)$ is the regularization term that avoids over-fitting, T denotes the total number of leaf nodes, W denotes the fraction of leaf nodes, γ and λ are the regularization parameters controlling the complexity of the model. To achieve fast optimization of the objective function, the loss function is subjected to a second-order Taylor expansion and can be approximated after omitting the constant term as:

$$obj^t\approx\sum\nolimits_{i=1}^{n}l\left(y_i,\widehat{y}_i^{(t-1)}+g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i)\right)+\Omega(f_t) \tag{8}$$

where g_i and h_i are the first and second order partial derivative of the loss function respectively. Next, all leaf nodes are reclassified through precise greedy algorithm is used to continuously select the single-leaf node with the largest gain. Then the objective function of the XGBoost model is shown in Eq. (9):

$$obj^t=-\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j}g_i\right)^2}{\sum_{i\in I_j}h_i+\lambda}+\gamma T \tag{9}$$

2.2. Python system for landslide susceptibility assessment (PSLSA)

Python is a simple but powerful open programming language that provides efficient data structures and effective object-oriented programming. The processing of geospatial data based on the Python language can be carried out in a sequential manner according to the programming steps without human intervention. The Python language is also widely used in machine learning, as it is embedded

with a large number of open source algorithm libraries. Therefore, this study uses the Python language to develop a regional landslide susceptibility assessment system. The processing flow of regional landslide susceptibility evaluation based on the data-driven approach mainly includes three stages: (1) collection and processing of landslide susceptibility influencing factors; (2) construction of landslide susceptibility evaluation models; and (3) susceptibility mapping and result evaluation. Correspondingly, the proposed PSLSA in this paper can also be divided into three processing modules (Fig. 1): module of influencing factors, module of landslide susceptibility

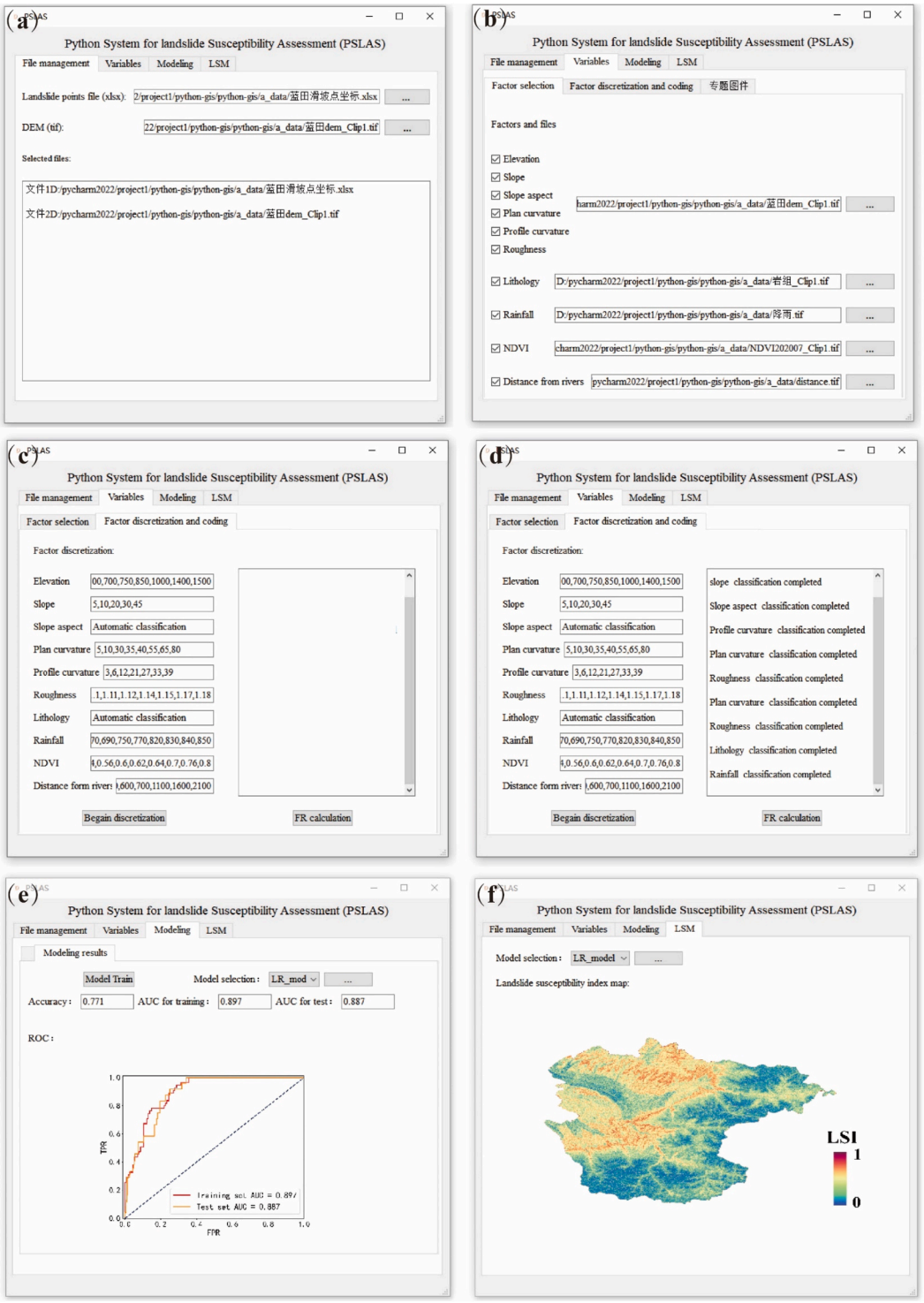


Fig. 1. The graphic user interface of the system for landslide susceptibility assessment.

models, and module of evaluation results.

2.2.1. Module of influencing factors

The main function of the evaluation factor module is to select influencing variables and reclassify the continuous variables aiming to facilitate analysis and simplify calculations. Firstly, in the file management interface (Fig. 1a), users need to input the digital elevation model (DEM) file of the study area and the EXCEL file containing the coordinates of the landslide points. In this part, the system extracts the existing recorded landslide raster (code with 1) and randomly selects an equal number of non-landslide raster (code with 0). Subsequently, based on previous research, we set a total of 10 influencing factors as input data (Fig. 1b), which can be divided into four major categories: topographic factors (elevation, slope, slope aspect, plan curvature, profile curvature, surface roughness), geological factors (lithology), hydrological factors (rainfall, distance to river) and land cover factors (normalized vegetation cover index (NDVI)). After importing the map for each factor, the user needs to set a set of thresholds to classify the continuous factors into sub-categories (Fig. 1c). The system automatically calculates the statistical relationship between each sub-category and historical landslides using the FR method (see Section 2.1 for procedure) (Fig. 1d). The theme maps are then converted by normalization into textual data that can be used directly for landslide susceptibility modeling task.

2.2.2. Module of landslide susceptibility modeling

Four ML models are available in this module: LR, MLP, SVM and XGBoost which are detailed described in Section 2.1. These models are initialized by introducing the scikit-learn and XGBoost libraries in Python. The modeling dataset was extracted based on the text data generated above containing the influencing factors and the targets of these factors (landslides or non-landslides). The dataset is also automatically divided into a training set and a test set, with 70 % of the training set and 30 % of the test set. In order to obtain a better fit and prediction performance of these models, in this part the user needs to initialize the hyperparameters of each model in the configuration file.

2.2.3. Module of modeling results

The main function of this module is to assess and compare the performance of the four ML models (Fig. 1e) according to the predicted landslide susceptibility index (Fig. 1f). The evaluation indicators include the receiver operating characteristic (ROC) curve and the accuracy value (Accuracy) obtained from the confusion matrix. The area under the curve (AUC) can also be used to evaluate the predictive performance of different models [37]. These metrics are divided into two groups for the training and testing datasets, and the robustness of each model can be assessed by comparing these indicators on different datasets.

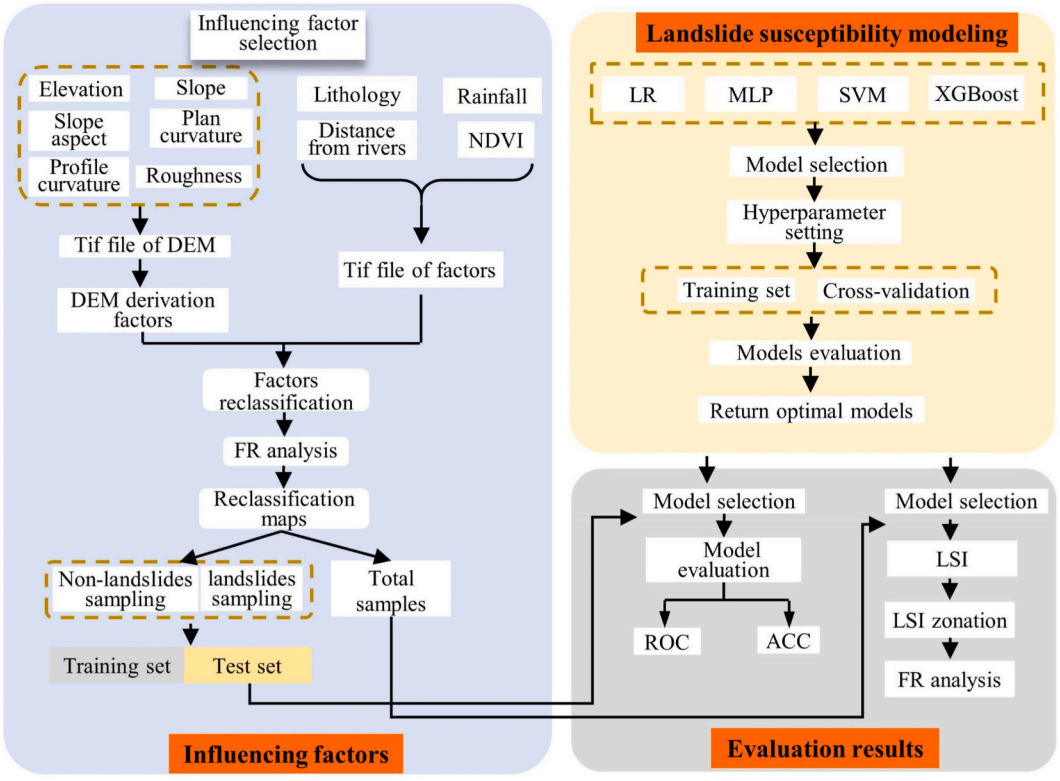


Fig. 2. The procedure of automatic landslide susceptibility assessment in this study.

2.3. General process of landslide susceptibility evaluation

When the proposed PSLAS produce was developed for regional landslide susceptibility assessment, it mainly included the six steps (Fig. 2). (1) The selection of landslide influencing factors that were required for susceptibility assessment, and imported related raster files. The topographic processing produce were imaged in PSLAS, since these related factors of elevation, slope, slope aspect, plane curvature, profile curvature and surface roughness could be autumnally calculated from the DEM. (2) The system read the factor maps and divided them into multiple sub-categories according to manually defined reclassification points. Subsequently, the FR analysis was implemented for each category and converted each theme map to text data for saving. (3) The selected factor maps were integrated into a complete variable dataset. Except for landslide inventory, the same number of non-landslide points were also selected randomly, both of them can provide the necessary information on factor conditions for whether landslide prone occurrence. (4) Aiming to each ML model, searching space for hyperparameter combinations was manually defined. Then PSLAS automatically invokes the dataset formed in the previous step for training models and hyperparameter optimization. (5) The landslide susceptibility modeling was carried out with the optimal hyperparameters solution, and accuracy (ACC) and AUC were used to evaluate the prediction precisely among different ML models. (6) Through comparison among the prediction precise, the appropriate model was selected for the landslide susceptibility assessment for the study area, and the landslide susceptibility index map can be viewed and derived for any model.

3. Study area and data processing

3.1. Study area

Lantian County of Shaanxi Province, China, was selected as the study area (Fig. 3a) to test the propose PSLAS. It lies between longitudes ranging from 109°07′E to 109°49′E and 33°50′ to 34°19′ N latitude, with a distance of 22 km from Xi'an city. The east-west length of the study area is approximately 64 km and the north-south width reaching 55 km, with a total area of 1977 km^2. The county of Lantian belongs to the piedmont basin between the Lishan and Qinling Mountains. The terrain of the region slopes from southeast to northwest, with the mountainous area occupying the majority. The southeastern part is characterized by the Qinling Mountain range, while the central and western parts consist of alternating valleys and terraces. Towards the north, loess hills form the predominant geographical feature [38]. Within the Lantian county, there are signiticant differences in elevation. Mount Wangshun claims the highest peak, rising to an approximate altitude of 2300 m, whereas the valley of the Ba River descends to a minimum elevation of 418 m.

The region falls within the temperate continental semi-humid zone, characterized by an average annual precipitation ranging from 600 to 900 mm. Through meticulous field investigations and extensive analysis of remote sensing imagery, a total of 79 landslides were identified throughout the study area (Fig. 3b). These landslides exhibit varying surface areas ranging from 200 m^2 to 10^6 m^2. The mean thickness of these landslides is estimated to be 5.2 m. In terms of their material composition, the most of them belong to loess landslides triggered by intense rainfall events.

3.2. Landslide influencing factors

As described in Section 2.2, the system developed in this study provides 10 common landslide susceptibility influencing factors as input choices. By comparison with previous studies in similar regions, all factors were found to be associated with loess landslide

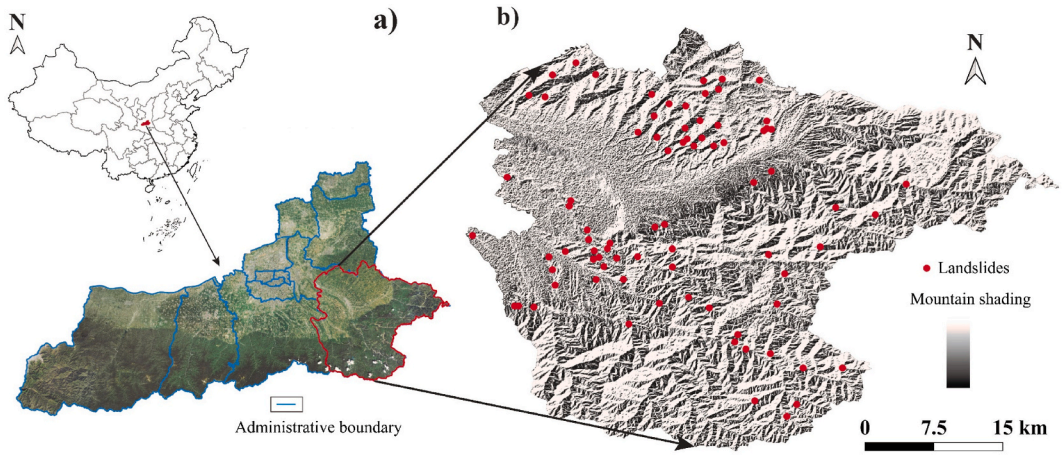


Fig. 3. Location of the study area and spatial distribution of the landslides : (a) location of Lantian County; (b) spatial distribution of the landslides in Lantian County.

susceptibility [39,40]. In order to make the defined subinterval classification of each factor more accurate and reliable, we first divided each factor into several small subcategories according to equal intervals, and calculated the FR value for each subcategory [41]. Subsequently, the subcategories with similar FR values were combined, while the FR value equal to 1 was also used as the threshold value for classification. Finally, the classification threshold values were obtained and set in the system. According to previous research [25], the number of sub-intervals between 5 and 12 is most appropriate for regional landslide susceptibility. Therefore, in the study, for continuous-type factors, the final number of reclassified subcategories all ranged from 5 to 12. The specific preparations for each factor and the process of FR analysis are described below:

Elevation (Fig. 4a): Elevation reflects the potential energy of the slope and controls the magnitude of stress values within the slope. Besides, different elevations affect the environment in which the slope is located, such as temperature and human engineering activities. The DEM raster with a resolution of 30 m was generated from an open-source data site (http://www.gscloud.cn/). The remaining influencing factors all have the same number of cells, shape and resolution of the DEM raster. The elevation range for the whole area is between 500 and 2500 m above sea level, with the highest number of landslides and FR-value at medium to high elevations, especially concreted on 600–1000 m. According to the FR analysis, the factor of elevation was divided into eight sub-intervals (Fig. 5a).

Slope (Fig. 4b): Slope reflects the steepness of the terrain and is one of the most important factors in describing topographic features and affecting the stability of slopes. The Slope was generated using a DEM filter in this system. The FR statistics showed that landslides were mainly distributed in the areas with slope ranging of 0–60°. Specifically, a total of over 70 % of landslides in the range of 10–30°. Based on the FR results, the slope of the whole area was divided into six sub-zones (Fig. 5b).

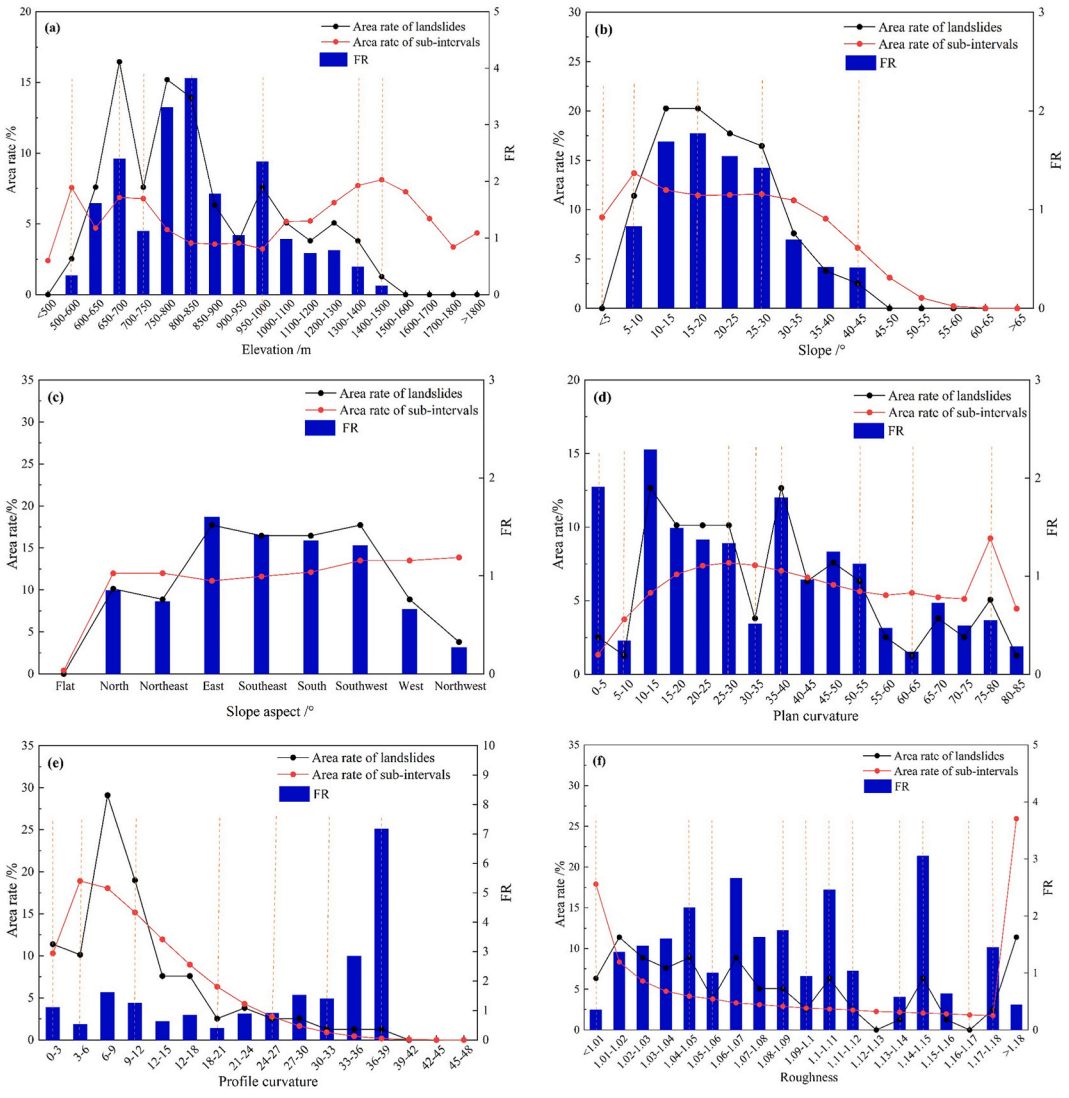


Fig. 4. Frequency ratio analysis and reclassification of the influencing factors in this study.

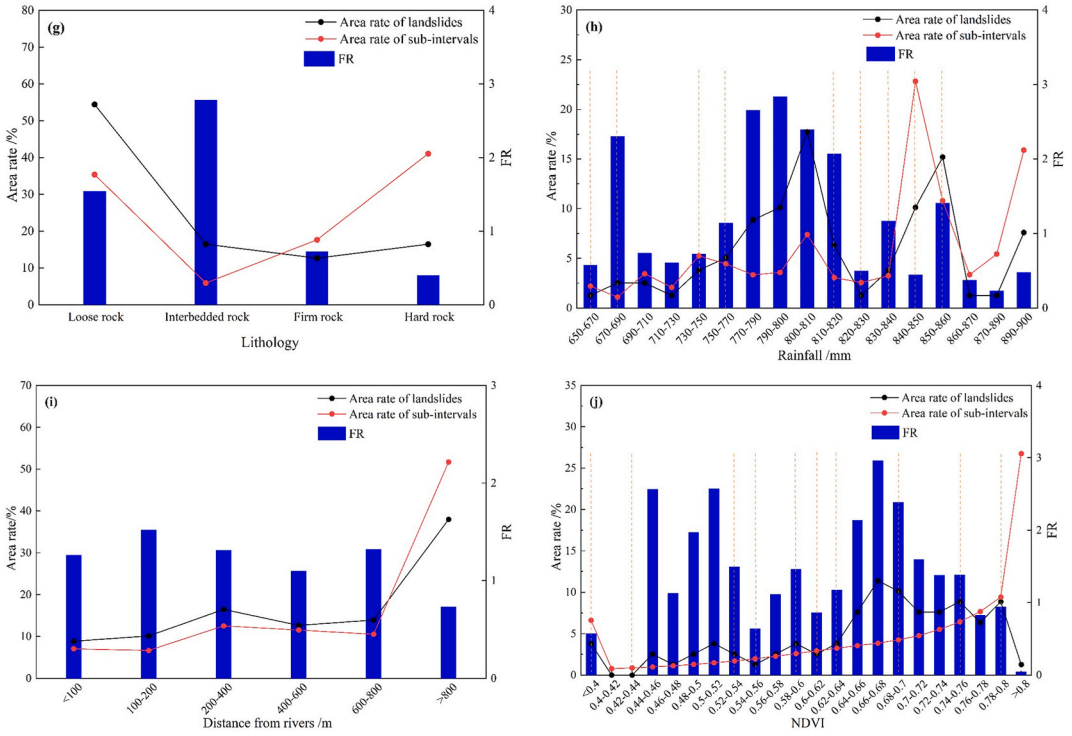


Fig. 4. (*continued*).

Slope aspect (Fig. 4c): Slope aspect can influence conditions such as temperature, insolation, vegetation cover and, in turn, hydrological processes and stability of slopes. In the literature, slope aspect is generally classified into nine sub-categories based on geographical orientation, namely flat (−1), north (337.5°–22.5°), northeast (22.5°–67.5°), east (67.5°–112.5°), southeast (112.5°–157.5°), south (157.5°–202.5°), southwest (202.5°–247.5°), west (247.5°–292.5°) and northwest (292.5°–337.5°) (Fig. 5c).

Plane curvature (Fig. 4d): The plan curvature can affect the ponding of water in the area and thus the stability of the slopes. Plan curvature values ranged from 0 to 90 across the region. The majority number of landslides and the largest FR values occurred between 10 and 40. The plan curvature was divided into nine sub-intervals based on the FR results, (Fig. 5d).

Profile curvature (Fig. 4e): Profile curvature measures the rate of change of the surface slope along the direction of maximum descent, thus indicating the characteristics of the terrain surface. Both plan curvature and profile curvature could be automatically generated directly in this the Python-based system of this study using the DEM map as input data. In the present study, this profile curvature was divided into eight sub-categories (Fig. 5e), 0–3, 3–6, 6–12, 12–21, 27–33, 33–39 and > 39. The statically results showed that historical landslides were mainly distributed in the areas with lower values of profile curvature.

Roughness (Fig. 4f): The factor of roughness reflects the degree of undulation of the surface. This factor is calculated from the slope file using the formula of $1/cos(slope)$, where a small value of surface roughness indicates a gentler surface, and. According to Fig. 5f, the surface roughness values for the whole area ranged from 0 to 1.2 and were eventually divided into 12 sub-categories. Areas with smaller roughness values seem to be accompanied by more landslides than areas with larger roughness values.

Lithology (Fig. 4g): Lithology can provide the material basis for landslide mass and is one of the most important geological factors [42]. This factor was derived from the 1:50,000 geological map of the whole study area. For the lithology map, it was divided into four types of engineering geological rock groups, in the order of hard rock, firm rock, interbedded rock and loose rock. Among of these types of rocks, the majority of historical landslides occurred on loose rocks, and the FR values of both loose and interbedded rocks were greater than 1, indicating that these two types of lithology were conducive to landslide formation (Fig. 5g).

Rainfall (Fig. 4h): Rainfall infiltration increases the weight of slopes and weakens the strength of geotechnical bodies, which is an important trigger for landslide hazards [43,44]. The average annual rainfall of several rainfall stations in Lantian County during 1999–2019 was collected to generate annual rainfall contours for the whole region. Subsequently, the inverse weight interpolation method was used to obtain the annual rainfall factor map. It can be found that the average annual rainfall values in the region range from 650 to 900 mm and that the majority of the landslides occur in the area with 770–820 mm of annual rainfall. The areas with an average annual rainfall of greater than 850 mm are mainly located in the northeastern part of the study area, where the bedrock overburden is thick. This limits the infiltration of rainfall, resulting in a relative low probability of landslides (Fig. 5h).

Distance from rivers (Fig. 4i): Rivers can erode the slopes and change the distribution of groundwater. Hence, the distance from rivers is considered to be an important influence factor influencing the occurrence of landslides. According to the scale of the river network in the study area, the whole area was divided into six buffer zones of 100 m, 200 m, 400 m, 600 m, 800 m, and more than 800

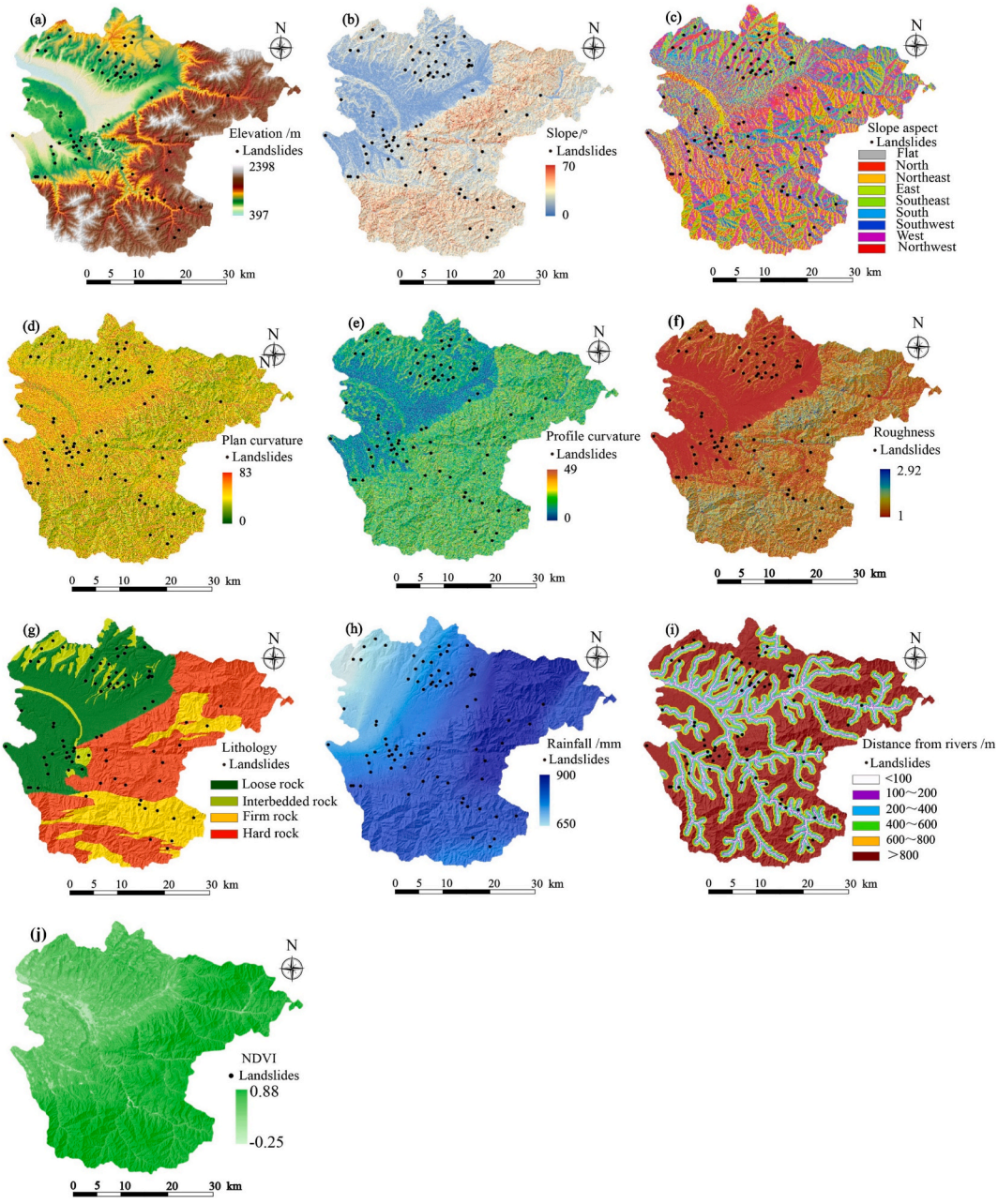


Fig. 5. The influencing factors used for landslide susceptibility assessment of Lantian County: (a)elevation; (b)slope; (c) aspect; (d) plan curvature; (e) profile curvature; (f) roughness; (g)lithology; (h) rainfall; (i) distance to river; (j) NDVI.

m, respectively (Fig. 5i).

NDVI (Fig. 4j): NDVI is a remotely sensed indicator reflecting the status of land cover vegetation, with values between −1 and 1. Higher values indicate denser vegetation cover. This factor (Fig. 5j)was calculated using Landsat8 OLI (30 m resolution) remote sensing data according to the following equation:

$$\mathrm{NDVI} = (\mathrm{NIR} - \mathrm{R}) / (\mathrm{NIR} + \mathrm{R}) \tag{10}$$

where NIR refers to the near infrared band in the image and *R* refers to the infrared band.

4. Results

4.1. Influencing factors processing results

As shown in the aforementioned process, the FR results of each influencing factor after reclassification are shown in Table 2. It can be found that the significant differences of FR values between the classified sub-intervals for each factor. This suggests that the distribution of landslides shows varying degrees of concentration or sparseness in different sub-regions.

As for the elevation factor, historical landslides were mainly distributed within the area of medium elevation, because the lower elevation was the plain area of Lantian County, while the areas with higher elevation were almost uninhabited. Hence, the impact of human activities on the stability of geological bodies was limited. Similarly, areas with moderate slope values also had the largest landslide FR, while slope angles were in the range of 0–5° and greater than 45° both have no recorded landslide. One on the one hand, areas with low slope values are not conducive to landslide formation, on the other hand, areas with high slopes which greater than 45° have a small total area of coverage. There are therefore no reports of landslide events in areas of low angle and high slope in the study area. For the rainfall factor, the average landslide density was significantly higher in areas with annual rainfall in the range of 840–900 mm than in areas with rainfall less than 820 mm. For the lithology factor, among the four types of engineering rocks, the FR values of landslides in areas with loose rock and interbedded rock were 1.54 and 2.78, respectively, indicating that such engineering geological conditions were prone to landslides. The FR values for landslides in areas with firm rock and hard rock were reduced to 0.72 and 0.40, indicating that landslides were not densely developed in these areas. Hence, these two rock groups were not conducive to landslide development. For NDVI, landslides were most densely distributed in the range of 0.62–0.76, indicating that landslides in the study area were more frequently distributed in areas with middle vegetation cover. The frequency of landslides was small where areas with density vegetation cover (NDVI ranges 0.76–0.88), and the FR value for these regions was less than 1. For the distance from the river factor, although nearly 40 % of the landslide points were located in the areas with distance from rivers greater than 800 m, but the FR values for these areas were less than 1. On the contrary, the limited reported landslide events in all other sub-areas with the distance from rivers lower than 800 m, but FR values for these areas were all greater than 1. The FR results indicated that landslides were more likely to occur in areas where close to the river banks.

Overall, these FR analysis results were consistent with the actual spatial distribution of landslides in the study area. Furthermore, the relationship between landslide points and the relative spatial location of the influencing factors suggested that the correlation between factors and the spatial distribution of landslides should focus not only on the number of landslides but also on the density of landslides. In addition, when the distribution of landslides on different sub-intervals of the same factor exhibits significant differences, it indicates a better reclassification effect.

In this study, ten landslide influencing factors were used to describe the likelihood of landslide occurrence. However, if there is a high degree of correlation between these factors, the reliability and interpretability of the model results can be seriously affected.

Table 2
The frequency ratio results after the reclassification of influencing factors.

Factors	subintervals	FR	Factors	subintervals	FR	Factors	subintervals	FR
Elevation	<600	0.25	Plan curvature	0–5	1.91	Surface roughness	<1.01	2.83
(m)	600–700	2.08		5–10	0.34		1.01–1.05	0.63
	700–750	1.12		10–30	1.58		1.05–1.06	1.00
	750–850	3.54		30–35	0.51		1.06–1.09	0.49
	850–1000	1.70		35–40	1.80		1.09–1.1	1.06
	1000–1400	0.72		40–55	1.11		1.1–1.11	0.41
	1400–1500	0.16		55–65	0.35		1.11–1.12	0.96
	>1500	0.00		65–80	0.58		1.12–1.14	3.51
Slope (°)	0–5	0.00		>80	0.28		1.14–1.15	0.33
	5–10	0.83	Profile curvature	0–3	1.11		1.15–1.17	3.02
	10–20	1.73		3–6	0.54		1.17–1.18	0.69
	20–30	1.48		6–12	1.45		>1.18	2.28
	30–45	0.53		12–21	0.65	Slope aspect	Falt (−1)	0.00
	>45	0.00		21–27	0.90	(°)	337.5°–22.5°	0.85
Rainfall (mm)	650–670	1.74		27–33	1.48		22.5°–67.5°	0.74
	670–690	0.43		33–39	4.08		67.5°–112.5°	1.60
	690–750	1.42		>39	0.00		112.5°–157.5°	1.42
	750–770	0.88	NDVI	<0.4	0.57		157.5°–202.5°	1.36
	770–820	0.40		0.4–0.44	0.00		202.5°–247.5°	1.31
	820–830	2.03		0.44–0.54	1.93		247.5°–292.5°	0.66
	830–840	0.86		0.54–0.56	0.64		292.5°–337.5°	0.27
	840–850	2.25		0.56–0.6	1.30	Distance from diver (m)	<100	1.26
	850–860	0.71		0.6–0.62	0.86		100–200	1.52
	860–900	2.44		0.62–0.64	1.17		200–400	1.31
Lithology	Hard rock	1.54		0.64–0.7	2.50		400–600	1.10
	Firm rock	2.78		0.7–0.76	1.44		600–800	1.32
	Interbedded rock	0.72		0.76–0.8	0.89		>800	0.73
	Loose rock	0.40		>0.8	0.05			

Therefore, in order to ensure that factors without significant multicollinearity are input into the model, we used the variance inflation factor (VIF) method [45] to diagnose the multicollinearity between the impact factors. Generally, the value of VIF ranges from 1 to 10. A large VIF value indicates an increase in multicollinearity. Whereas, when the VIF value is greater than 10, it indicates a serious multicollinearity problem. In this study, all ten influencing factors used had VIF values ranges of 1–2, indicating only low multicollinearity (Fig. 6). Therefore, in this study, all ten influencing factors were accepted for landslide susceptibility assessment.

4.2. *Landslide susceptibility assessment results and comparative analysis*

The optimal hyperparameter configurations for each model (LR, MLP, SVM and XGBoost) were obtained by setting up the search grid and cross-validation methods the specific hyperparameters for the four models [46] were shown in Table 3. The system of PLSAS then uses the optimal modeling solution to predict landslide susceptibility in the study area. The landslide susceptibility results for Lantian County generated by PLSAS were shown in Fig. 7. Each landslide susceptibility map (LSM) generated by the various ML models was divided into five susceptibility levels using natural breakpoints method: very high (VH), high (H), medium (M), low (L) and very low (VL). The distribution of landslide susceptibility for the four models (LR (Fig. 7a), SVM (Fig. 7b), MLP (Fig. 7c) and XGBoost (Fig. 7d)) follows the similar general trend, but there were still different spatial distribution patterns. For example, the LSM generated by XGBoost model identified more areas of VH susceptibility in the northern part of the study area. Whereas the MLP model had recognized more areas of VH susceptibility in the southern part of the study site compared to the other models. The LSM generated by LR model had a slightly smaller area of VH susceptibility in the northwest compared to the other three maps.

In order to reveal more clearly the differences in the spatial distribution patterns of various susceptibility maps, the percentage of area occupied by each landslide susceptibility level and the percentage of landslides were calculated (Table 4). These calculated results showed that the VL and L susceptibility areas for each LSM were larger than the other landslide susceptibility levels. The accumulation area of these two low levels for each map accounted for 57.9 % (LR), 54.8 % (SVM), 62.1 % (MLP) and 51.7 % (XGBoost) of the total area, respectively. This suggested that even conducted on the different modeling models, a common conclusion that could be drawn was that the overall landslide susceptibility of the whole study area is at a low level. In addition, the recorded historical landslides falling within these two susceptibility levels was limited, accounting for only 8.9 % (LR), 8.9 % (SVM), 7.6 % (MLP) and 7.6 % (XGBoost), respectively. On the contrary, the areas labeled with VH and H susceptibility levels accounted relative smaller compared to other areas labeled with different susceptibility levels. For each map, the percentage of areas zoned in VH and H levels were 26.6 % (LR), 27.5 % (SVM), 25.3 % (MLP) and 31.9 % (XGBoost), respectively. Whereas these areas in each map identified 74.7 % (LR), 81.0 % (SVM), 83.6 % (MLP) and 78.5 % (XGBoost) of historical landslides. This indicates that all four models were better predictors as they were able to identify the majority of historical landslides in smaller areas of high susceptibility level.

Furthermore, the FR method was applied to analysis the relationship between the historical landslides and the susceptibility zonation for each LSM (Table 4). A clear finding is that as the level of landslide susceptibility increases, the FR value increases significantly for these four maps. The FR values for the VL and L susceptibility zones in the four maps were all less than 1, while the FR values for the VH and H susceptibility zones were all greater than 1. The results of FR analysis indicated that there were significant differences in landslide densities within different susceptibility levels. The areas with higher susceptibility level being associated with higher landslide densities and landslides being more likely to occur in the area. The FR results can also verify the effectiveness of the four built-in ML models in this system.

From the perspective of model comparison, the SVM and MLP models identified more landslides in the VH susceptibility zonation and had higher landslide density. Thus, the mapping performance of SVM and LR were considered better than the maps generated by LR and XGBoost models. Besides, the FR values in the susceptibility zonation of M for the LR model was slightly greater than 1 (1.06), while these values for the other three models all less than 1 (0.57 for SVM, 0.7 for MLP, and 0.8 for XGBoost for 0.8), indicating that the LR model performs slightly worse than several other models.

We then have counted the statistical distribution of landslide susceptibility index (LSI) obtained from the four different models (LR

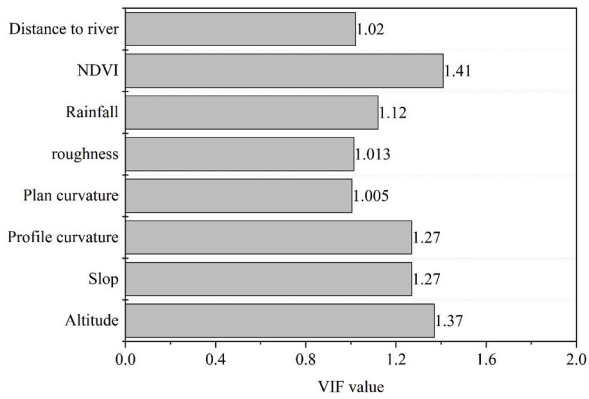


Fig. 6. VIF results for the ten influencing factors.

Table 3
The optimal hyperparameters for LR, MLP, SVM and XGBoost.

Models	Hyperparameters
LR	Penalty = "l2", C = 1.0, Max_iteration = 300, Slover = "lbfgs"
MLP	Hidden layer size = (24,8), Activation function = "relu", Slover = "sgd", Alpha = 0.7
SVM	C = 1.3, Kernel = "rbf", Gamma = 0.29, tol = 0.001
XGBoost	n_estimators = 90, max_depth = 5, subsample = 0.8, colsample_bytree = 0.7, gamma = 0.1, min_child_weight = 5

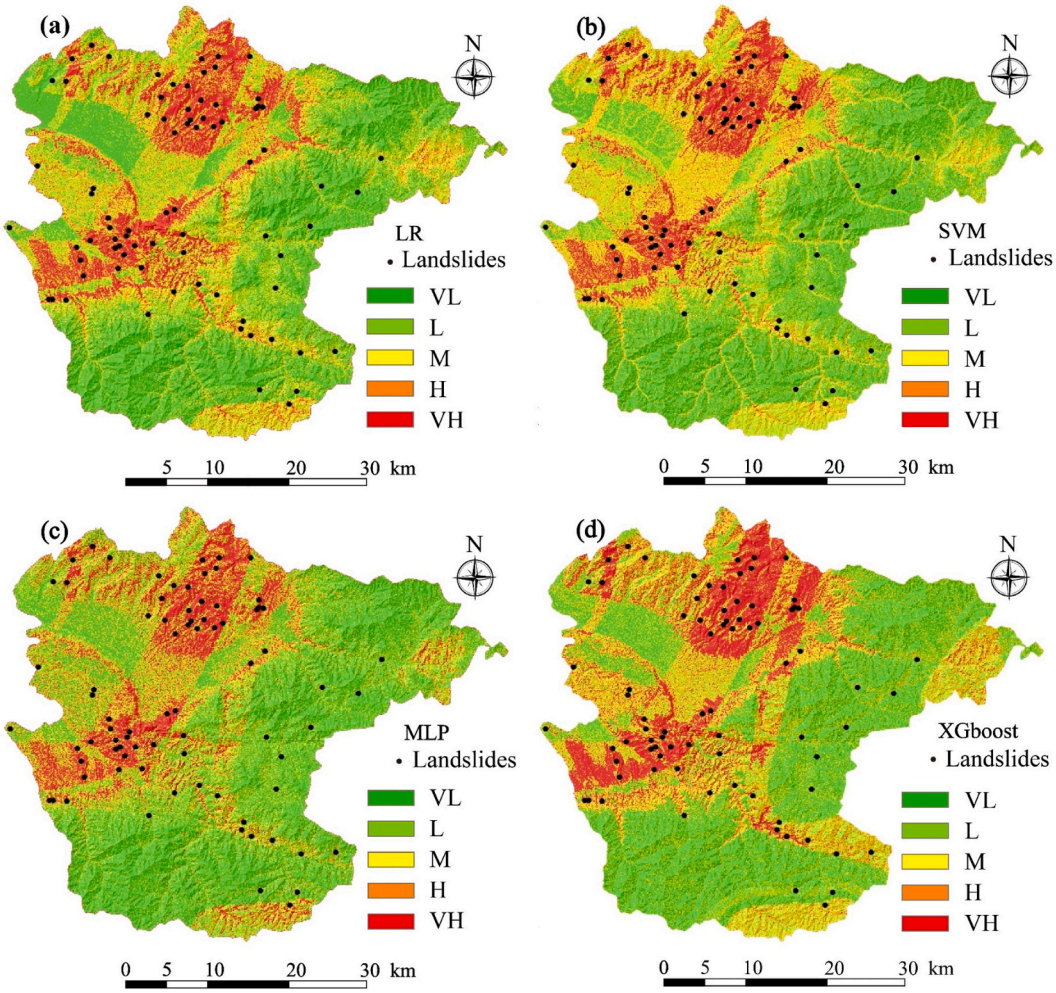


Fig. 7. Landslide susceptibility zonation of Lantian County: (a) LR model; (b) SVM model; (c) MLP model; (d) XGBoost model.

(Fig. 8a), SVM (Fig. 8b), MLP (Fig. 8c) and XGBoost (Fig. 8d)). Besides, the mean value and standard deviation (SD) of LSIs were used to reflect the overall mean and degree of dispersion. As seen in Fig. 8, the distribution patterns of the susceptibility indices predicted by the four models showed significant differences. The LSIs of LR and MLP models showed peaks in the range less than 0.1, but these values of LR model showed a decreasing or slightly increasing trend around the maximum susceptibility value (i.e. susceptibility index of 1). On the contrary, the LSIs of MLP model showed an increasing trend around the susceptibility index of 1. The LSIs of SVM model showed a roughly normal skewed distribution. The LSI for most computation cells ranged between 0.15 and 0.17 and the number of cells decreased outwards on both sides of the peak LSI. The LSIs calculated by XGBoost model also had a peak of the number of cells as the LSI in a smaller range of from 0.05 to 0.16, but when the LSI was greater than 0.2, the number of cells tend to be relatively evenly distributed. The four models were ranked from smallest to largest in terms of the mean value of the LSI was MLP (0.193) < LR (0.209) < XGBoost (0.224) < SVM (0.388); the rank of standard deviation was: SVM (0.168) < XGBoost (0.272) < LR (0.28) < MLP (0.29).

The probability distribution functions of the above mentioned LSI were then calculated using the Matlab code, as shown in Fig. 9. The results of probability distribution of LSI were similar to those revealed by the above statistical indicators. The peak values of LSIs

Table 4
The distribution pattern and frequency ratio of landslides in the landslide susceptibility maps obtained from different models.

Models	Areas of landslide susceptibility levels and landslide statistics. (A: Rate of area in landslide susceptibility levels; B: Rate of landslides in landslide susceptibility levels; C: FR value of landslide susceptibility levels)				
	VL	L	M	H	VH
LR	A = 35.7 %; B = 3.8 %; C = 0.11	A = 22.2 %; B = 5.1 %; C = 0.23	A = 15.6 %; B = 16.5 %; C = 1.06	A = 12.5 %; B = 20.3 %; C = 1.63	A = 14.1 %; B = 54.4 %; C = 3.87
SVM	A = 31.6 %; B = 1.3 %; C = 0.4	A = 23.2 %; B = 7.6 %; C = 0.33	A = 17.7 %; B = 10.1 %; C = 0.57	A = 16.0 %; B = 22.8 %; C = 1.43	A = 11.5 %; B = 58.2 %; C = 5.09
MLP	A = 42.9 %; B = 2.5 %; C = 0.06	A = 19.2 %; B = 5.1 %; C = 0.26	A = 12.6 %; B = 8.9 %; C = 0.70	A = 11.0 %; B = 20.3 %; C = 1.84	A = 14.3 %; B = 63.3 %; C = 4.42
XGBoost	A = 34.1 %; B = 1.3 %; C = 0.04	A = 17.6 %; B = 6.3 %; C = 0.36	A = 17.4 %; B = 13.9 %; C = 0.80	A = 16.6 %; B = 24.1 %; C = 1.54	A = 15.3 %; B = 54.4 %; C = 3.56

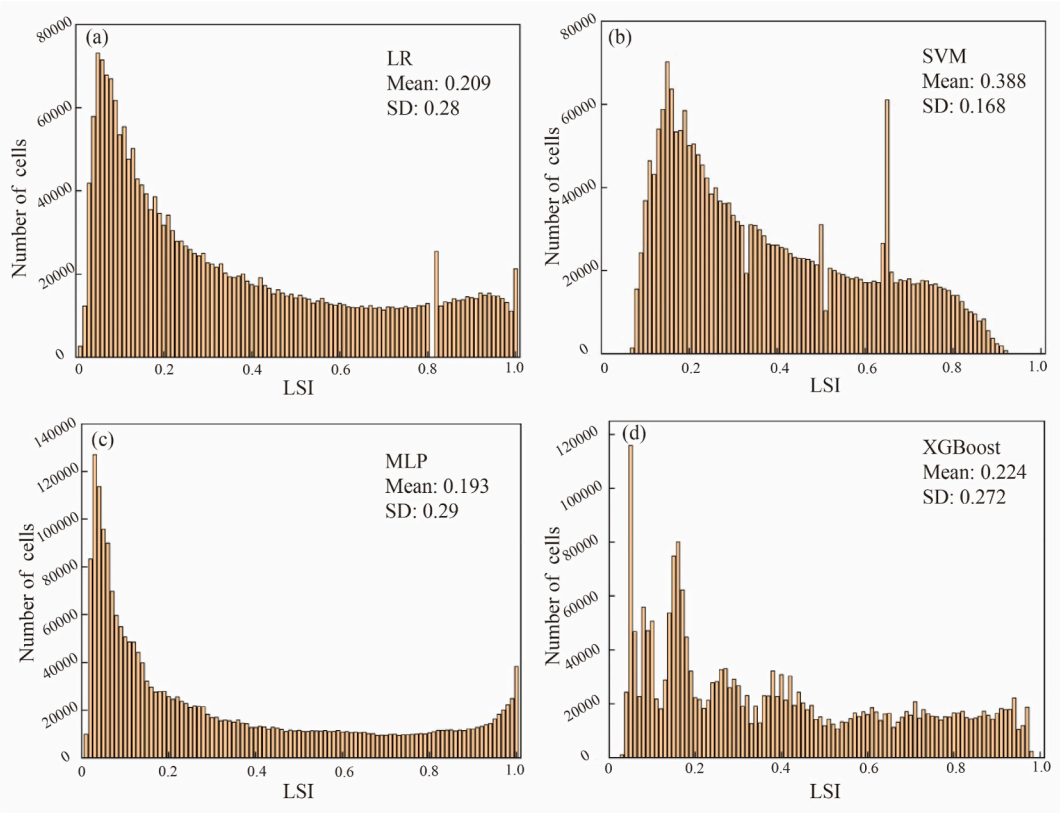


Fig. 8. Interval of susceptibility index and number of pixels for different models: (a) LR; (b) SVM; (c) MLP; (d) XGBoost.

calculated by LR, MLP and XGBoost models have smaller values than the SVM model. In addition, SVM model had the largest mean value and a smaller degree of dispersion (SD), therefore it can be found that SVM model had a more concentrated distribution of LSIs. The SVM model also has a smaller number and density of cells at the two poles of LSIs (probability density values equal to 0 for LSI <0.05 and >0.96), which explains why the SVM model had a smaller area of VL and VH susceptibility zones than the other models (Table 4). However, the SVM identified the second largest proportion of landslides in the VH susceptibility level which was only lower than the MLP model. But the LSM generated by SVM had the largest FR value, indicating that the SVM predicted historical landslides well and identified more historical landslides with only a smaller number of VH susceptibility cells.

As described in Section 2.2, the landslide susceptibility assessment system can automatically calculate ROC curves for different models to quantitatively evaluate and compare model performance. Fig. 10 shows the AUC accuracy for the training (Fig. 10a) and test datasets (Fig. 10b) for the four models. For the training set, the AUC values for the four models were 99.7 % (MLP), 91.9 % (SVM), 87.6 % (LR) and 86.7 % (XGBoost) respectively; while the prediction accuracies for test set were 88.2 % (SVM), 83.8 % (LR), 81.2 % (XGBoost) and 80.9 % (MLP) respectively. It can be found that the accuracies of all the models were greater than 80 % and some of them were greater than 90 %. In addition, the difference of prediction accuracies between the test set and the training set were not significant. These results showed that the Python system (PSLSA) for landslide susceptibility assessment were promising and the four

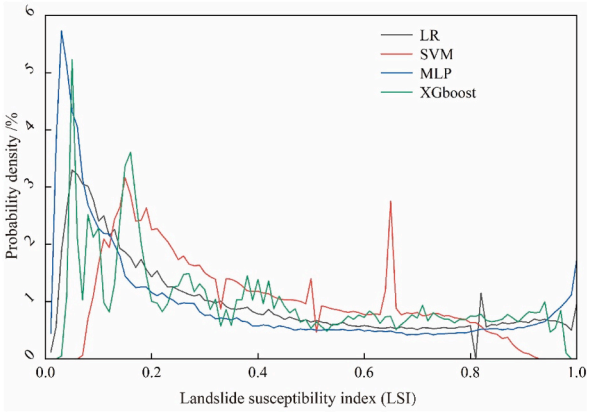


Fig. 9. Probability distribution function of susceptibility index of different models.

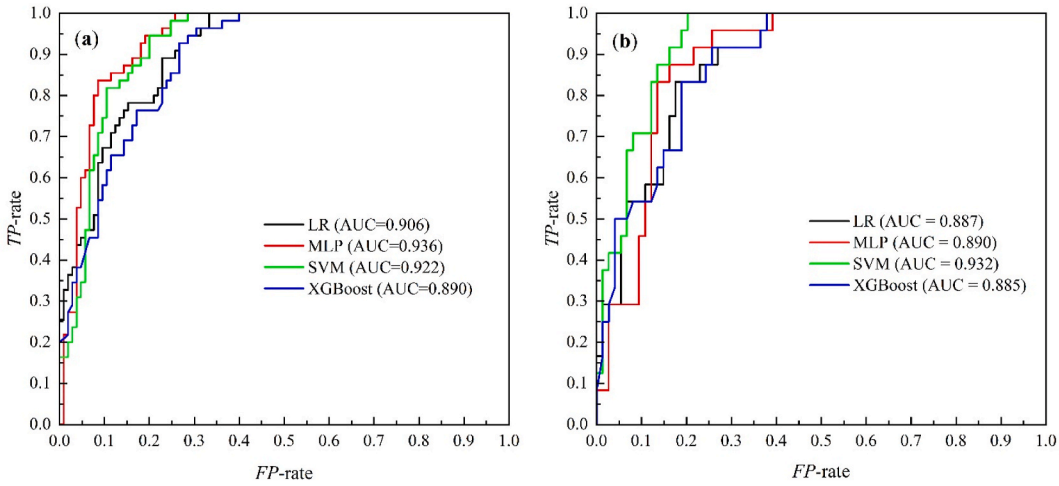


Fig. 10. ROC curves of the four models: (a) result for training dataset, (b) result for testing dataset.

ML models implied in this system were very robust. From the perspective of model performance comparison, the MLP model had the highest fitting accuracy on training set, but the worst accuracy on test set, indicating a weak generalization capability for MLP. The SVM model had the second highest accuracy on training set after the MLP and the highest prediction accuracy on test set. Comparative results shown that SVM was the most suitable for landslide susceptibility assessment in Lantian County. In addition, unlike the AUC values, which was calculated through a confusion matrix, this system also calculated the classification accuracy (ACC) value for each model. The final ACC value for each model were 0.771 (SVM), 0.75 (LR), 0.708 (XGBoost) and 0.708 (MLP). Among of them, the SVM still being the best performing model.

5. Discussion

5.1. Insights on the similar ROC values but different LSI distribution

In this study, four machine learning models (LR, MLP, SVM and XGBoost) were employed to evaluate the landslide susceptibility. These models were also been optimized to identify the optimal parameters and configuration, which were then used to obtain the best solution for predicting landslide susceptibility. Based on the modeling results, we reconstructed the LSI values for each cell in the study area to obtain the statistical information, including the density distribution, mean and SD values of LSIs for each model. Although all of these models reached relatively high accuracy in predicting landslide occurrence, with no significant difference for ROC values (ranging from 0.88 to 0.93 on the test set). However, the LSI maps (Fig. 7) and statistical results revealed different distribution patterns. One possible reason for this is that the machine learning approach in the training set was modeled with the aim of a binary classification task. The goal of modeling was to distinguish between landslide and non-landslide points, not the LSI value of each cell. Hence, the ROC index used to evaluate the accuracy of the final results tends to achieve satisfactory accuracy. However, for the intermediate results in the process of modeling, different machine learning models obtain differential LSI distributions due to different

learning strategies. The uncertainty regarding LSI further affects the results of the landslide susceptibility zonation maps. In addition, this highlights the limitations of using a single metric to assess model performance, as it may not fully capture the local conditions of certain geomorphological features [11].

5.2. Advantage and limitations of the proposed procedures

Conventional regional landslide susceptibility assessment is usually a multidisciplinary and integrated analysis process where the application of GIS, statistical tools and programming (for modelling calculations) are necessary. However, such multidisciplinary applications can be technically difficult or time-consuming for the general user [12,24]. The Python system developed in this study integrates these multiple processes of landslide susceptibility assessment, including the geographic data processing, susceptibility modelling, and evaluation of results. In this system, only a small number of parameters and operations need to be defined and operated by the user throughout the landslide susceptibility assessment process, which does not require much experience. Once the required influencing factors and landslide inventory obtained, the landslide susceptibility assessment and accuracy calculation can be carried out automatically in an efficient and accurate process. The system is now open source on Github (https://github.com/GuoGroup-EngineeringGeology/PSLSA.git, PSLSA is the Python based System for Landslide Susceptibility Assessment), which is very user-friendly. The results of the practical application in Lantian County, Shaanxi Province, showed that all integrated ML models have an accuracy rate of over 80 % on both the training and test sets, with some models exceeding 90 % accuracy, in line with engineering practice. In addition, although the system was carried out in the Loess Plateau, China, it can be generalized to other similar environments as it contains general input data for landslide susceptibility assessment.

However, the system needs to be improved in certain fields. Firstly, the limited selection of influencing factors. The system currently provides 10 fixed influencing factors as input which is commonly used for landslide susceptibility assessment. Although current inputs covered many aspects that effecting the stability of slopes, there is still a lack of task- or environment-specific considerations. As a result, the proposed landslide susceptibility assessment system is limited in performing some specific tasks. For example, human engineering activities have direct impact on slope stability, which may have a greater impact in areas prone to landslides due to mountain cutting and large-scale hydropower construction. But this factor was not considered in this system. It should to be noted that there is no optimal combination for influencing factors selection that can be used to assess landslide susceptibility, since the effects of different combinations are dependent on the characteristics of the study area. Therefore, one of our future tasks is to provide a wider choice of influencing factors and to offer more flexible factor selection options in this system. However, it is up to the user to decide whether to use a particular factor or not, depending on the geological setting and landslide characteristics of the study area. Another aspect of the landslide susceptibility assessment system that needs to be improved is the limited diversity of modeling methods. In the literature, there is no consensus suggestions on which models are more appropriate for landslide susceptibility assessment. Although this system provides four benchmark ML models for users to choose, they still have limitations. Deep learning models currently offer better performance than ML methods in many fields, and some scholars are beginning to apply such models to landslide susceptibility assessment [13,47,48]. Therefore, embedding more ML models and deep learning models will also be an important task in the development of this system. When the choice of influencing factors and evaluation models is more extensive, the system will be able to be applied to more similar studies in other study areas.

6. Conclusion

(1) This study developed a landslide susceptibility evaluation system (PSLAS) with integrated machine learning models based on the Python language. With the required of based landslide susceptibility influencing factors, the user was able to automate the landslide susceptibility modeling and assessment using four built-in models. The application case in Lantian County showed that the proportion of historical landslide points identified by this system in VH and H susceptibility levels reached 50%–78 %, and the FR values were all greater than 1. These evaluation results indicate that the system can accurately perform the tasks of evaluating landslide susceptibility on a regional scale.

(2) The spatial distribution of the susceptibility index obtained from the four modeling methods used in this study is generally similar, but there are still detailed differences between different susceptibility maps. For each model, the fitting accuracies on training set were 99.7 % (MLP), 91.9 % (SVM), 87.6 % (LR) and 86.7 % (XGBoost), and prediction performance on test set were 88.2 % (SVM), 83.8 % (LR), 81.2 % (XGBoost) and 80.9 % (MLP). This suggests that comparing the performance of different modeling methods within the same study area is necessary to select an appropriate modeling approach. For region of Lantian County, the SVM model may be a better choice for conducting effective landslide susceptibility evaluation, since it is not only more accurate but also more robust (the accuracy of the training set and the accuracy of the test set were similar).

(3) The system developed in this study can be used to automatically evaluate regional landslide susceptibility independently of the GIS platform. It avoids the complex data pre-processing and subsequent accuracy analysis processes. Moreover, the code of this system has been open-sourced, making it more user-friendly for ordinary users lacking expertise in geological engineering. Hence, it holds great application prospects and potential.

Data availability statement

Data associated with this study will be made available on request.

CRediT authorship contribution statement

Zizheng Guo: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Fei Guo:** Conceptualization, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. **Yu Zhang:** Formal analysis, Investigation, Resources, Software, Writing – review & editing. **Jun He:** Data curation, Investigation, Validation, Visualization, Writing – original draft. **Guangming Li:** Formal analysis, Funding acquisition, Investigation, Project administration, Supervision, Writing – review & editing. **Yufei Yang:** Formal analysis, Investigation, Visualization, Writing – original draft. **Xiaobo Zhang:** Formal analysis, Methodology, Software, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research is funded by National Natural Science Foundation of China (No.42107489, 42307248), Natural Science Foundation of Hebei Province (D2022202005), the Open Fund of Hubei Key Laboratory of Disaster Prevention and Mitigation (China Three Gorges University) (No.2022KJZ02), the Open Fund of Badong National Observation and Research Station of Geohazards (No.BNORSG-202304), the Natural Science Foundation of Hubei Province (No.2022CFB557), the Open Fund of Key Laboratory of Geological Hazards on Three Gorges Reservoir Area (China Three Gorges University) of Ministry of Education (No.2022KDZ14), the 111 Project of Hubei Province (Grant Number 2021EJD026), and the Tianjin Municipal Bureau of Planning and Natural Resources Projece (Grant Number 2022–40).

References

[1] D. Petley, Global patterns of loss of life from landslides, Geology 40 (2012) 927–930.
[2] M.J. Froude, D.N. Petley, Global fatal landslide occurrence from 2004 to 2016, Nat. Hazards Earth Syst. Sci. 18 (2018) 2161–2181.
[3] Z. Guo, L. Chen, K. Yin, D.P. Shrestha, L. Zhang, Quantitative risk assessment of slow-moving landslides from the viewpoint of decision-making: a case study of the Three Gorges Reservoir in China, Eng. Geol. 273 (2020), 105667.
[4] A. Su, M. Feng, S. Dong, Z. Zou, J. Wang, Improved statically solvable slice method for slope stability analysis, Journal of Earth Science 33 (2022) 1190–1203.
[5] Z. Chang, F. Catani, F. Huang, G. Liu, S. Meena, J. Huang, C. Zhou, Landslide susceptibility prediction using slope unit-based machine learning models considering the heterogeneity of conditioning factors, J. Rock Mech. Geotech. Eng. 15 (5) (2023) 1127–1143.
[6] Z. Guo, J.V. Ferrer, M. Hürlimann, V. Medina, C. Puig-Polo, K. Yin, D. Huang, Shallow landslide susceptibility assessment under future climate and land cover changes: a case study from southwest China, Geosci. Front. 14 (4) (2023), 101542.
[7] B. Pradhan, A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS, Comput. Geosci. 51 (2013) 350–365.
[8] Q. Li, D. Huang, S. Pei, J. Qian, M. Wang, Using physical model experiments for hazards assessment of rainfall-induced debris landslides, Journal of Earth Science 32 (5) (2021) 1113–1128.
[9] V. Medina, M. Hurlimann, Z. Guo, A. Lloret, Fast physically-based model for rainfall-induced landslide susceptibility assessment at regional scale, Catena 201 (2021), 105213.
[10] L. Dong, H. Zhu, F. Yan, S. Bi, Risk field of rock instability using microseismic monitoringdata in deep mining, Sensors 23 (3) (2023) 1300.
[11] P. Reichenbach, M. Rossi, B.D. Malamud, M. Mihi, A review of statistically-based landslide susceptibility models, Earth Sci. Rev. 180 (2018) 60–91.
[12] J. Goetz, A. Brenning, H. Petschko, P. Leopold, Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling, Comput. Geosci. 81 (2015) 1–11.
[13] Y. Wang, Z.C. Fang, M. Wang, L. Peng, H. Hong, Comparative study of landslide susceptibility mapping with different recurrent neural networks, Comput. Geosci. 138 (2020), 104445.
[14] H. Pourghasemi, B. Pradhan, C. Gokceoglu, M. Mohammadi, H. Moradi, Application of weights-of-evidence and certainty factor models and their comparison in landslide susceptibility mapping at Haraz watershed, Iran, Arabian J. Geosci. 6 (7) (2013) 2351–2365.
[15] D. Bui, B. Pradhan, O. Lofman, I. Revhaug, Landslide susceptibility assessment in vietnam using support vector machines, decision tree, and naïve Bayes models, Math. Probl Eng. (2012), 974638.
[16] A. Merghadi, A. Yunus, J. Dou, J. Whitelry, B. ThaiPham, D. Bui, R. Avtar, B. Abderrahmance, Machine learning methods for landslide susceptibility studies: a comparative overview of algorithm performance, Earth Sci. Rev. 207 (2020), 103225.
[17] Y.A. Nanehkaran, T. Pusatli, J. Chengyong, J. Chen, A. Cemiloglu, M. Azarafza, R. Derakhshani, Application of machine learning techniques for the estimation of the safety factor in slope stability analysis, Water 14 (22) (2022) 3743.
[18] Z. Guo, B. Tian, G. Li, D. Huang, T. Zeng, J. He, D. Song, Landslide susceptibility mapping in the Loess Plateau of northwest China using three data-driven techniques-a case study from middle Yellow River catchment, Front. Earth Sci. 10 (2023), 1033085.
[19] W. Chen, Y. Li, GIS-based evaluation of landslide susceptibility using hybrid computational intelligence models, Catena 195 (2020), 104777.
[20] X. Tang, T. Machimura, W. Liu, J. Li, H. Hong, A novel index to evaluate discretization methods: a case study of flood susceptibility assessment based on random forest, Geosci. Front. 12 (6) (2021), 101253.
[21] P. Kainthura, N. Sharma, Machine learning driven landslide susceptibility prediction for the Uttarkashi region of Uttarakhand in India, Georisk 16 (3) (2022) 570–583.
[22] X.Y. Shao, S.Y. Ma, C. Xu, Q. Zhou, Effects of sampling intensity and non-slide/slide sample ratio on the occurrence probability of coseismic landslides, Geomorphology 363 (2020), 107222.
[23] N.R. Andabili, M. Safaripour, Identification of precipitation trend and landslide susceptibility analysis in Miandoab County using MATLAB, Environ. Monit. Assess. 197 (7) (2022) 472.
[24] J. Torizin, N. Schussler, M. Fuchs, Landslide Susceptibility Assessment Tools v1.0.0b – Project Manager Suite: a new modular toolkit for landslide susceptibility assessment, Geosci. Model Dev. (GMD) 15 (7) (2022) 2791–2812.
[25] F. Huang, Z. Cao, J. Guo, S. Jiang, S. Li, Z. Guo, Comparisons of heuristic, general statistical and machine learning models for landslide susceptibility prediction and mapping, Catena 191 (2020), 104580.

[26] S. Bartolini, A. Cappello, J. Marti, C. Del Negro, QVAST: a new Quantum GIS plugin for estimating volcanic susceptibility, Nat. Hazards Earth Syst. Sci. 13 (11) (2013) 3031–3042.
[27] M. Alvioli, I. Marchesini, P. Reichenbach, M. Rossi, F. Ardizzone, F. Fiorucci, F. Guzzetti, Automatic delineation of geomorphological slope units with r. slopeunits v1.0 and their optimization for landslide susceptibility modeling, Geosci. Model Dev. (GMD) 9 (11) (2016) 3975–3991.
[28] E. Sahin, I. Colkesen, S. Acmali, A. Akgun, A. Aydinoglu, Developing comprehensive geocomputation tools for landslide susceptibility mapping: LSM tool pack, Comput. Geosci. 144 (2020), 104592.
[29] Z. Guo, O. Torra, M. Hurlimann, C. Abanco, V. Medina, FSLAM: a QGIS plugin for fast regional susceptibility assessment of rainfall-induced landslides, Environ. Model. Software 150 (2022), 105354.
[30] A. Aditian, T. Kubota, Y. Shinohara, Comparison of GIS-based landslide susceptibility models using frequency ratio, logistic regression, and artificial neural network in a tertiary region of Ambon, Indonesia, Geomorphology 318 (2018) 101–111.
[31] H. Hong, H.R. Pourghasemi, Z.S. Pourtaghi, Landslide susceptibility assessment in Lianhua County (China): a comparison between a random forest data mining technique and bivariate and multivariate statistical models, Geomorphology 259 (2016) 105–118.
[32] C. Xu, X.W. Xu, F.C. Dai, A.K. Saraf, Comparison of different models for susceptibility mapping of earthquake triggered landslides related with the 2008 Wenchuan earthquake in China, Comput. Geosci. 46 (2012) 317–329.
[33] S. Lee, B. Pradhan, Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models, Landslides 4 (1) (2007) 33–41.
[34] D. Li, F. Huang, L. Yan, Z. Cao, J. Chen, Z. Ye, Landslide susceptibility prediction using particle-swarm-optimized multilayer perceptron: comparisons with multilayer-perceptron-only, BP neural network, and information value models, Appl. Sci. 9 (18) (2019) 3664.
[35] H. Wei, W. Shu, L. Dong, Z. Huang, D. Sun, A waveform image method for discriminating micro-seismic events and blasts in underground mines, Sensors 20 (15) (2020) 4322.
[36] T.Q. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, USA, 2016, pp. 785–794.
[37] M. Hürlimann, Z. Guo, C. Puig-Polo, V. Medina, Impacts of future climate and land cover changes on landslide susceptibility: regional scale modelling in the Val d' Aran region (Pyrenees, Spain), Landslides 19 (2022) 99–118.
[38] Y. Wu, Z.Y. Zhu, Z.G. Rao, S.F. Qiu, T. Yang, Mid-Late Quaternary loess-paleosol sequence in Lantian's Yushan, China: an environmental magnetism approach and its paleoclimatic significance, Chin. Sci. Bull. 55 (26) (2010) 2989–3000.
[39] Y. Tang, F. Feng, Z. Guo, W. Feng, Z. Li, J. Wang, Q. Sun, H. Ma, Y. Li, Integrating principal component analysis with statistically-based models for analysis of causal factors and landslide susceptibility mapping: a comparative study from the loess plateau area in Shanxi (China), J. Clean. Prod. 277 (2020), 124159.
[40] Y. Tang, Z. Guo, L. Wu, B. Hong, W. Feng, X. Su, Z. Li, Y. Zhu, Assessing debris flow risk at a catchment scale for an economic decision based on the LiDAR DEM and numerical simulation, Front. Earth Sci. 10 (2022), 821735.
[41] H. Shu, Z. Guo, S. Qi, D. Song, H.R. Pourghasemi, J. Ma, Integrating landslide typology with weighted frequency ratio model for landslide susceptibility mapping: a case study from lanzhou city of northwestern China, Rem. Sens. 13 (18) (2021) 3623.
[42] Y. Huang, C. Xu, X. Zhang, C. Xue, S. Wang, An updated database and spatial distribution of landslides triggered by the milin, tibet Mw6.4 Earthquake of 18 november 2017, Journal of Earth Science 32 (5) (2021) 1069–1078.
[43] C. Dai, W. Li, D. Wang, H. Lu, Q. Xu, J. Jian, Active landslide detection based on sentinel-1 data and InSAR Technology in zhouqu county, gansu province, northwest China, Journal of Earth Science 32 (5) (2021) 1092–1103.
[44] Z. Guo, L. Chen, L. Gui, J. Du, K. Yin, H.M. Do, Landslide displacement prediction based on variational mode decomposition and WA-GWO-BP model, Landslides 17 (2020) 567–583.
[45] S. Nikoobakht, M. Azarafza, H. Akgün, R. Derakhshani, Landslide susceptibility assessment by using convolutional neural network, Appl. Sci. 12 (12) (2022) 5992.
[46] Y. Li, S. Abdallah, On hyperparameter optimization of machine learning algorithms: theory and practice, Neurocomputing 415 (2020) 295–316.
[47] M. Azarafza, M. Azarafza, H. Akgün, P.M. Atkinson, R. Derakhshani, Deep learning-based landslide susceptibility mapping, Sci. Rep. 11 (1) (2021), 24112.
[48] L. Zhu, L. Huang, L. Fan, J. Huang, F. Huang, J. Chen, Z. Zhang, Y. Wang, Landslide susceptibility prediction modeling based on remote sensing and a novel deep learning algorithm of a cascade-parallel recurrent neural network, Sensors 20 (2020) 1576.