Journal of Rock Mechanics and Geotechnical Engineering 16 (2024) 877–894

Contents lists available at ScienceDirect

Journal of Rock Mechanics and Geotechnical Engineering

journal homepage: www.jrmge.cn

Full Length Article

How do the landslide and non-landslide sampling strategies impact landslide susceptibility assessment? — A catchment-scale case study from China

Zizheng Guo [a,b], Bixia Tian [a], Yuhang Zhu [c,*], Jun He [a], Taili Zhang [d]

[a] School of Civil and Transportation Engineering, Hebei University of Technology, Tianjin, 300401, China
[b] Hebei Key Laboratory of Earthquake Disaster Prevention and Risk Assessment, Sanhe, 065201, China
[c] Faculty of Engineering, China University of Geosciences, Wuhan, 430074, China
[d] Nanjing Center, China Geological Survey, Ministry of Natural Resources, Nanjing, 210016, China

ARTICLE INFO

Article history:
Received 13 January 2023
Received in revised form
10 May 2023
Accepted 9 July 2023
Available online 12 January 2024

Keywords:
Landslide susceptibility
Sampling strategy
Machine learning
Random forest
China

ABSTRACT

The aim of this study is to investigate the impacts of the sampling strategy of landslide and non-landslide on the performance of landslide susceptibility assessment (LSA). The study area is the Feiyun catchment in Wenzhou City, Southeast China. Two types of landslides samples, combined with seven non-landslide sampling strategies, resulted in a total of 14 scenarios. The corresponding landslide susceptibility map (LSM) for each scenario was generated using the random forest model. The receiver operating characteristic (ROC) curve and statistical indicators were calculated and used to assess the impact of the dataset sampling strategy. The results showed that higher accuracies were achieved when using the landslide core as positive samples, combined with non-landslide sampling from the very low zone or buffer zone. The results reveal the influence of landslide and non-landslide sampling strategies on the accuracy of LSA, which provides a reference for subsequent researchers aiming to obtain a more reasonable LSM.

1. Introduction

Landslides are one of the major geological disasters all over the world, causing considerable human casualties and economic losses every year (Petley, 2012). For example, the southeastern part of China often suffers from rainstorm-triggered landslide events in summer, which may lead to dozens of deaths and millions in economic losses (Su et al., 2015; Zhao et al., 2019; Ma et al., 2022a; Guo et al., 2023a). Thus, spotting the areas exposed to landslides is an important task for land planning and regional security (Chen and Li, 2020; Guo et al., 2020a). Recently, landslide susceptibility mapping (LSM) has been a commonly used tool for achieving this goal, and extensive stakeholders have begun to assess and reduce landslide risk through it (Brenning, 2005; van Westen et al., 2008; Agterberg, 2022).

Landslide susceptibility assessment (LSA) is to obtain the spatial distribution of landslide occurrence probability in a region by analyzing the correlation among historical landslides and environmental factors (Fell et al., 2008; Pradhan, 2010). A normal LSA procedure mainly includes the sampling of landslides (positive) and non-landslide (negative) datasets, and the determination of influencing factors, landslide susceptibility modelling and mapping, and accuracy analysis of results (Barik et al., 2017). Among them, input data commonly differs in terms of landslide and non-landslide samples, leading to a significant source of uncertainty in LSA. Specifically, it is important to accurately express the spatial shape of landslide samples and the spatial location of non-landslide samples because they can affect the nonlinear correlation between samples and factors during modelling, which further impacts the accuracy of LSA (Arnone et al., 2016).

The landslide locations are usually determined from historical landslides, remote sensing images, and field investigations (Guzzetti et al., 2012; Guo et al., 2020b; Smith et al., 2021). Then the landslides should be digitized for the analysis of the relationship between landslides and environmental factors (Peng et al., 2014; Paryani et al., 2021). At present, expression forms of landslide samples in the literature mainly include points (Guo et al., 2022),

* Corresponding author.
E-mail address: cugzyh@cug.edu.cn (Y. Zhu).
Peer review under responsibility of Institute of Rock and Soil Mechanics, Chinese Academy of Sciences.

https://doi.org/10.1016/j.jrmge.2023.07.026

circles (Huang et al., 2022) or irregular shapes (dustpan shapes, semicircles, elongated bars, etc) (Galli et al., 2008; Moosavi et al., 2014). Specifically, the selection of landslide digital form largely depends on the study scale and landslide size. For studies at small scales (1:100,000 or smaller), it is commonly an operational challenge to accurately capture the actual boundary of landslides thus the forms of point and circle are mainly adopted under this condition (van Westen et al., 2006). Users only need to record limited but necessary information on landslides (e.g., location, type, and time) in the inventory. At medium (1:10,000–1:50,000) or large scales (>1:50,000), more information can be revealed in the landslide inventory since the test area is smaller, allowing for more detailed investigations, including initiation area, accumulation area, size and abundance. It is also common to identify the actual boundaries (irregular shapes) of landslides, especially with the development of remote sensing and drone techniques (Azarafza et al., 2018; Nikoobakht et al., 2022; Su et al., 2022). Considering the systematic error when it comes to the form of landslide points and circles, more studies prefer to digitalize the landslides as irregular shapes for the LSA (e.g., Shahabi and Hashim, 2015; Ma et al., 2022b), which align more closely with realistic situations. As Huang et al. (2022) pointed out, landslide samples expressed by irregular shapes (actual boundary) often have a higher result accuracy and a lower uncertainty in landslide susceptibility modelling. When a landslide inventory with irregular shapes is used for LSA, a data transformation from vector to raster (pixel) is required. However, when landslides are represented in GIS by pixels, their boundaries and pixel boundaries often do not completely overlap. There is not yet a unified standard determining whether a pixel crossing a landslide boundary should be considered part of the landslide samples. Although this aspect may significantly impact the outcome of the susceptibility analysis, the issue of the proper representation of landslide locations has not been widely discussed yet.

Regarding the sampling strategy of non-landslide dataset, there are mainly four kinds of ways: (i) sampling randomly from landslide-free areas, which is the most commonly used one in the literature (e.g. Okalp and Akgün, 2016; Bueechi et al., 2019; Azarafza et al., 2021), (ii) sampling from the buffer zone by defining a minimum distance between the landslide area and a landslide-free area (Xi et al., 2022), (iii) sampling from the low susceptibility zone obtained by constraining certain external factors, such as self-organizing neural network (Huang et al., 2017), similarity-based approach (Zhu et al., 2019), and (iv) sampling from terrain area with low slope angles or plain regions (Kavzoglu et al., 2014; Lucchese et al., 2021; Okalp and Akgün, 2022). However, all these strategies have inherent drawbacks which can cause some important aspects of uncertainty to the LSA (Zhu et al., 2019). For example, the negative samples generated by (i) and (ii) may be located on steep slopes that are susceptible to landslides. The strategy (iv) may identify the pixels near rivers which are also possibly prone to landslide occurrence. Hence, it is necessary to compare and assess the performance of different sampling strategies of non-landslide dataset in the LSA. Unfortunately, limited efforts have been made on this issue. Xi et al. (2022) compared the results from two scenarios, namely a set of negative samples created using different buffer distances (strategy (ii)) and a set of negative samples created by the Newmark-based method (strategy (iii)). Except this, more studies simply propose a specific method for non-landslide sampling and test its performance. The uncertainty stemming from this aspect remains due to the lack of comprehensive comparisons among different scenarios.

The commonly-used models for LSA can be divided into expert-based models (Sezer et al., 2017), physically-based models (Medina et al., 2021) and data-driven models (Eker et al., 2015; Zêzere et al., 2017). Generally, data-driven models can be classified as statistically-based and machine learning models (Pham et al., 2016; Reichenbach et al., 2018; Guo et al., 2023b). Compared with mathematical statistics, machine learning models have advantages in reflecting the nonlinear corrections between landslides and factors (Bueechi et al., 2019). Thus, models like Random Forest (RF), C5.0 Decision Trees (C5.0 DT), Logistic Regression (LR), Bayes Network (BN), Artificial Neural Networks (ANN), Multilayer Perceptron (MLP) and Support Vector Machine (SVM) have been widely developed and applied for LSA over the past decade (Bui et al., 2012; Conforti et al., 2014; Hong et al., 2020; Merghadi et al., 2020). Additionally, some recent advances in data processing techniques show that ensemble learning methods can further improve the performance of machine learning methods and alleviate their limitations (e.g. Bui et al., 2019; Chen and Li, 2020). However, it should be noted that a sole type of machine learning model may be coincidental regarding high accuracy, so it is necessary to consider different models to obtain more stable and reliable results. In this study, we used three models including C5.0 model, SVM LR models to generate landslide susceptibility maps, which herein we refer to as pre-LSM. The non-landslide dataset was sampled from the very low susceptibility zones in these pre-LSMs, which was subsequently used to generate final LSM. Considering the high accuracy and maturity of the RF model (Catani et al., 2013), this model was selected for the final landslide susceptibility modelling.

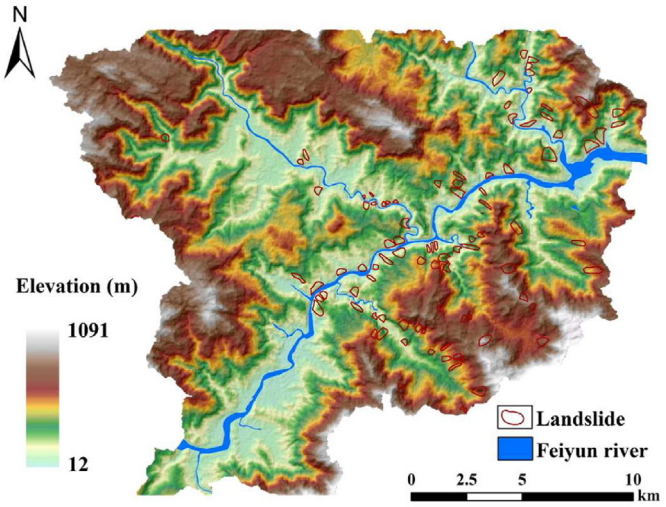

Fig. 1. Distribution of the landslides in the study area where the DEM with a resolution of 30 m is used as the base map.

In general, the application of machine learning in the field of LSA is relatively mature. However, there are imperfections in the modelling process. Specifically, there is limited research on the gridding method for landslide datasets and the selection of non-landslide samples. The main objective of the present study is to clarify the uncertainties from landslide and non-landslide sampling strategy and reveal their impacts on LSA. As far as we know, very few studies have discussed the combined effect of both aspects before. The Feiyun catchment in Zhejiang Province of Southeast China was taken as the study area. More specifically, our aims include: (i) design of landslide and non-landslide sampling strategies based on various sampling areas; (ii) constructing the combined scenarios of sampling strategies and conducting regional LSA based on machine learning models; and (iii) comparative analysis of predictive ability under different scenarios and effects of sampling strategies.

2. Materials

2.1. Study area

The present study was conducted in the Feiyun catchment, which is located southwest of Wenzhou, Zhejiang Province, China. It covers a total area of 388.17 km^2 and is geographically between latitudes 119°59′32″E and 120°15′30″E, longitudes 27°39′12″N and 27°50′29″N (Fig. 1). The landscape of the region is dominated by mountains and hills, and the elevation ranges from 12 m above sea level at the river valley to 1091 m at the highest peak, with the topography characterized by higher elevation in the northwest and lower elevations in the southeast. The area is a subtropical marine monsoon climate zone with a multi-year average temperature of 14 °C–18.5 °C and a multi-year average rainfall of approximately 1884.7 mm. The river systems in the area are well developed, with the Feiyun River being the main stream which flows through the catchment from west to east, and the settlements are mainly located along the river banks. The predominant strata outcrops in the region comprise volcanic rock, granites, Paleozoic clastic rock, and carbonate rock (Su et al., 2015; Wang et al., 2020).

2.2. Landslide inventory

The LSA by applying data-driven models is commonly based on an important assumption that future landslides will likely occur in areas with environments similar to those of historical landslides (Zêzere et al., 2017). Therefore, landslide inventory as a basis for understanding the spatial distribution of historical landslides in a region is essential (Huang et al., 2017). In this study, based on historical landslide reports, visual interpretation of remote sensing images and field geological surveys, 96 landslide points are identified in the inventory (Fig. 1). Their boundaries were also identified and subsequently digitized as polygons. Covering a total area of 8.8 km^2, these landslides constitute 0.23% of the total study area. These landslides are mainly developed along the river banks of the central and northwestern part, and identified as shallow type according to Varnes classification system (Varnes, 1978; Hungr et al., 2014). The most important triggering factor for them is heavy rainfall in rainy season, which commonly happens with typhoon events.

2.3. Data source and preparation of influencing factors

The data used in this study mainly includes the digital elevation model (DEM), satellite images, the survey map of the study area and water system. Detailed information and the purpose of the data are shown in Table 1.

The development and mechanism of landslides over a large area are complex and diverse, and mainly influenced by five types of factors, namely topography, hydrology, land cover, geology and others (e.g. human engineering activities) (Reichenbach et al., 2018; Goyes-Penafiel and Hernandez-Rojas, 2021). There is no agreement on the best combination of influencing factors for LSA so far. In this study, ten factors (Fig. 2) were selected for analysis, based on previous literature and overall characteristics of landslides in the region. It should be stated that the rainfall was not taken into account in this study because it was considered relevant to temporal probability of landslides, which was beyond the concept of landslide susceptibility in the widely accepted criterion (Fell et al., 2008).

Table 1
The information of data used in this study.

Type	Sources	Form	Propose
DEM	ASTER satellite	Raster	Preparing factor maps: Fig. 3a–f
Satellite image	Sentinel-2	Raster	Preparing factor map: Fig. 3h
Land use map	National Earth System Science Data Center	Vector	Preparing factor map: Fig. 3i
Soil type map		Vector	Preparing factor map: Fig. 3j
Water system	DEM	Vector	Preparing factor map: Fig. 3g

Topographic factors include aspect, elevation, slope, topographic position index (TPI), profile curvature and relief. Among them, slope is considered as the most important topographic factor that can directly control slope stability (Zhou et al., 2016). Aspect and TPI orientation indirectly affect landslide development by influencing the vegetation distribution and illumination of the slope (He et al., 2019). Elevation reflects the changes of the temperature, humidity and biodiversity on the slope (Shahri et al., 2019). Relief and profile curvature characterize the undulation of the terrain and subsequently control the water flow on the slope surface. The degree of relief was obtained by searching the 3 m × 3 m neighborhood grid unit and calculating its maximum elevation difference. The other factors were all generated by digital elevation model (DEM) data with spatial resolution of 30 m in GIS 10.6 software.

The hydrological factor was represented by the distance from the river, because it can affect groundwater level, drainage capacity of slopes and erosion on the bank slopes. Therefore, with the drainage line as the buffer center and 100 m intervals from 0 to 600 m, six equally spaced buffer zones were determined. Combined with the areas with the distance to river beyond 600 m, a total of seven levels were generated to represent the impact range of the river on the bank slopes.

Land cover influences the infiltration, drainage and anti-weathering ability of slopes, thus causing instability (Dao et al., 2020). The land cover data were downloaded from National Earth System Science Data Center (https://www.geodata.cn). It was divided into eleven categories, namely, dense forest, open forest, garden, shrubs, dense grass, open grass, paddy field, dry land, bare land, urban area and water, which were remarked from 1 to 11 in order. Moreover, normalized difference vegetation index (NDVI) was also considered, which can reflect the development of vegetation. Vegetation roots can reinforce soil and slow down soil erosion. Compared with densely vegetated areas, bare slopes without vegetation cover are commonly more prone to deformation when exposed to external conditions such as rainfall (Hürlimann et al., 2022). This map was calculated by using Sentinel-2 images at different wavelengths.

Geological factors are crucial in controlling landslide materials. Different soil types have various mechanical and hydrological properties, which have significant effects on occurrence of shallow landslides. Since most landslides in the study area were identified as shallow types, soil type was considered as geological factor. The soil type factor was obtained from National Earth System Science Data Center (https://www.geodata.cn), and it was divided into six types including red soil, yellow soil, paddy soil, submergic paddy soil, acid soil and brown soil, which were remarked from 1 to 6 in order. According to Chinese Soil Taxonomy, different soil types

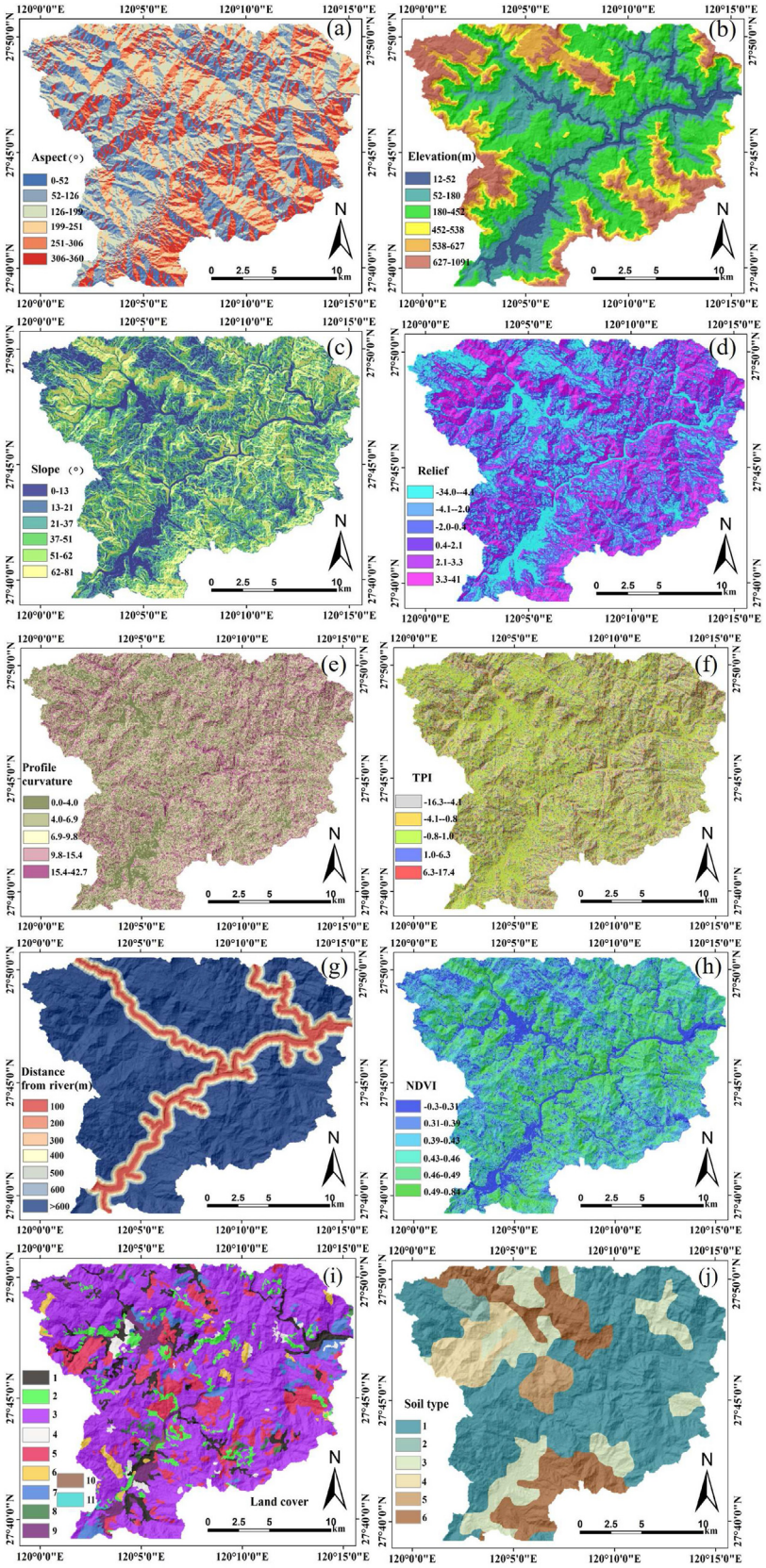

Fig. 2. The influencing factors used for: (a) Aspect, (b) Elevation, (c) Slope, (d) Relief, (e) Profile curvature, (f) TPI, (g) Distance from river, (h) NDVI, (i) Land cover, and (j) Soil type.

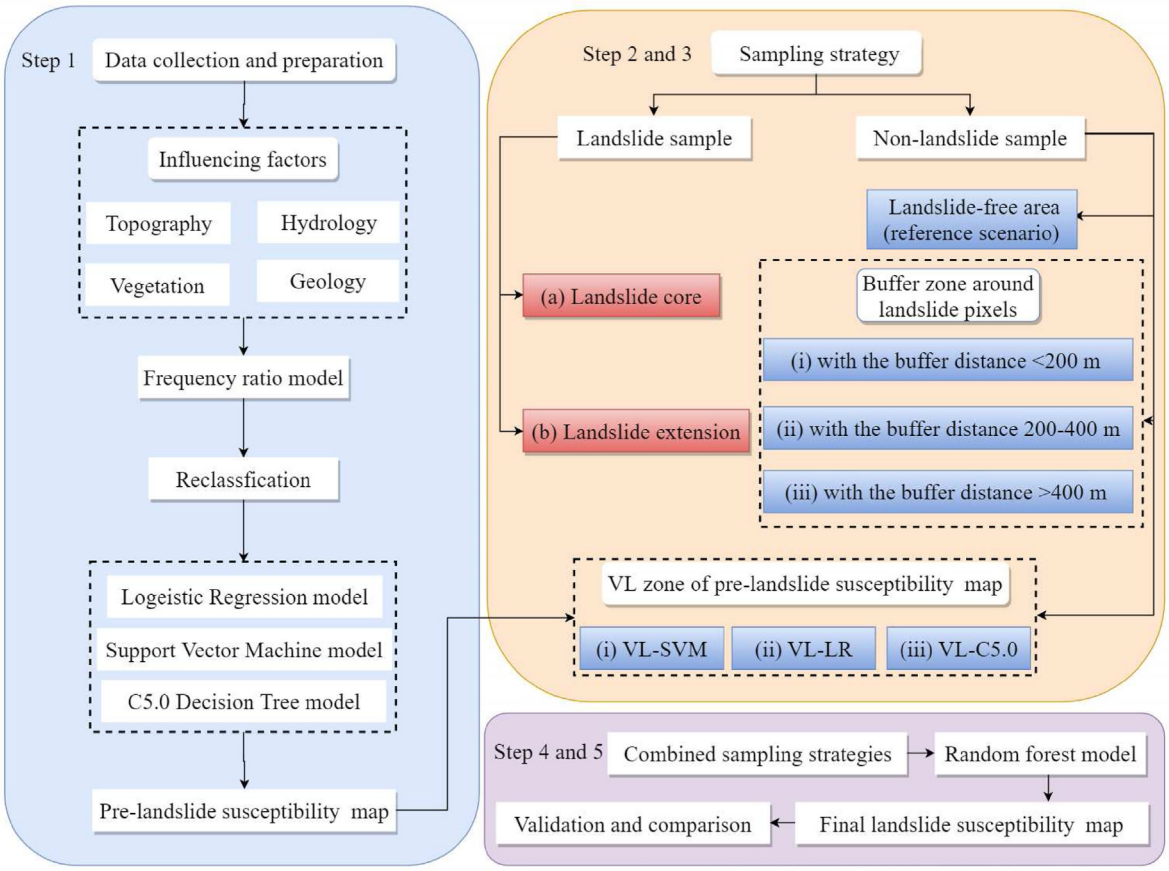

Fig. 3. The methodological framework of this study.

present different colors due to the difference of soil textures, so the colors can be used to represent soil type (Liu et al., 2020).

3. Methodologies

3.1. Modelling procedure

The flow chart of this study is shown in Fig. 3. Overall, there are totally five main steps as follows: (i) Data preparation and factor reclassification: All influencing factors were generated into raster with a resolution of 30 m using ArcGIS. Among these influencing factors, seven are continuous variables except for distance from river, land cover and soil type, which are discrete. On one side, it is time-consuming if every single value of the variables is used as the input parameter for the model without the division of classes. On the other side, a relevant source of subjectivity and uncertainty is introduced when splitting the input parameters into fixed classes with a certain of break values. To resolve this issue, the continuous influencing factors were first equally classified into 20 sub-categories, and then the frequency ratio of each classification was calculated in this study. Finally, the sub-categories with similar frequency ratios were grouped to form the final influencing factor classification. It should be mentioned that the final arbitrary number of classes was determined from 5 to 7, which fits with previous studies (Dou et al., 2020; Huang et al., 2020). The details of frequency ratio analysis for each factor will be described in section 4.1: (ii) Landslide sampling strategy: Two landslide expression styles were designed in this part. One is the rasters only within the landslide boundary (herein we call it by “landslide core”) were considered as the landslide dataset after rasterization of the landslide surface. The other one is that the rasters including landslide core and the rasters covering the landslide boundary were integrated as the entire group (herein we call it by “landslide extension”) (see the “sampling strategy” section for the detailed definitions of landslide core and landslide extension). (iii) Non-landslide sampling strategy: Seven scenarios were determined to sample non-landslide datasets, which can be divided into three categories. The first one was sampling non-landslides from landslide-free areas, which was commonly used in literature (Ali et al., 2022; Mehrabi, 2022). The second one was to obtain non-landslide samples from the buffer zone (within the distance of 200 m, 200–400 m, and larger than 400 m buffer zone to landslide boundary, respectively). The third method was to sample non-landslides from the very low susceptibility zones identified by some models in advance (herein we call it pre-LSM). Three machine models were applied to create pre-LSMs in this study, namely SVM, C5.0 and LR models. (iv) LSM: A total of fourteen scenarios were determined by combining the landslide and non-landslide sampling strategies mentioned above, and the LSM under each scenario was generated based on the RF model. (v) Validation and comparison: The receiver operating characteristic (ROC) curve and area under curve (AUC) value for each scenario were calculated to analyze the model performance and evaluate the effect of sampling strategy on LSA. It should be stated that three different machine learning models were employed in step (iii) to mitigate the chance of incidental errors in VL zone produced by a single model. During step (iv), the purpose of the application of the RF model instead of the three models mentioned above was to keep the sampling of landslide/non-landslide as the only variable in the test. The details on the model principles used will be described in continuation.

3.2. Frequency ratio (FR) model

The frequency ratio (FR) method is a traditional statistics model for regional LSA. The FR values of one factor greater than 1 indicate a positive relationship with landslide occurrence, and conversely FR values less than 1 represent a negative relationship (Shirzadi et al., 2017). It is now commonly used for the grading of influencing factors:

$$FR_{ij} = \frac{L_{ij}/TL}{C_{ij}/TC} \tag{1}$$

where i and j indicate the number and grading of impact factors respectively, L_{ij} indicates the area of landslides within class j of the i th influencing factor and TL represents the total area of landslides; C_{ij} indicates the area of class j of the i th impact factor and TC represents the total area of the study area.

3.3. Machine learning models

3.3.1. C5.0 decision tree (C5.0 DT)

The C5.0 Decision Tree model is based on a multistage or hierarchical decision structure. The core of the C5.0 DT model is using the rate of decline in information gain ratio (GR) as the basis for determining the optimal branching variables and segmentation thresholds. The GR can be expressed as following (Guo et al., 2021):

$$GainRatio = \frac{Gains(D,T)}{Ent(T)} \tag{2}$$

where D is the dataset and T is the predictor variable. The *Gains (D, T)* represents the entropy difference between the original and child nodes, which is calculated as below:

$$Gains(D,T) = \left[\sum_{i}^{n} P(C_i|D)\log_2 P(C_i|D)\right] \times \left(\sum_{j}^{m}\frac{|T_j|}{|D|}-1\right) \tag{3}$$

where C is the target variable, n is the category number of C, C_i ($i = 1, 2, \ldots, n$). The category number of T is m, T_j ($j = 1, 2, \ldots, m$). In this study, Adaboosting algorithm and Pruning was adopted to improve the generalization ability of the C5.0 DT model during the modelling process.

3.3.2. Support vector machine (SVM)

The SVM model is a sparse and robust classifier that uses a hinge loss function to compute empirical risk and adds a regularization term to the solution system to optimize structural risk. The SVM model is one of the common kernel learning methods for nonlinear classification by the kernel method (Zhang et al., 2019). Generally, a high-dimensional feature space was map data through nonlinear kernels and classified by separating hyperplane. The separating hyperplane can be defined as following:

$$\omega x + b = 0 \tag{4}$$

with the optimal solution to:

$$\min \frac{\|\omega\|^2}{2} + c\sum_{i=1}^{l}\xi_i \tag{5}$$

$$s.t.y_i(\omega \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \tag{6}$$

where ω determines the direction of the hyperplane, b is the bias, C represent the penalty, $x_i \in R^n$, $y_i \in \{-1, 1\}$, ξ_i is the slack factor used for classifier and the number of support vectors is l. The classification function is calculated as Eq. (7) by introducing kernel functions:

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{l} a_i y_i K(x_i \cdot x) + b\right) \tag{7}$$

where n is the number of the samples, and a_i is Lagrange multiplier.

3.3.3. Logistic regression (LR)

The logistic regression method is a classic linear regression model that has been widely applied to solve dichotomous problems in LSA (Kavzoglu et al., 2014). The main calculation formula of it can be written as following:

$$\mathrm{logit}(z) = k + a_1 y_1 + a_2 y_2 + a_3 y_3 + \cdots + a_n y_n \tag{8}$$

where z is the linear predictor for landslide, y_n is the characteristic value of influencing factors, k is a constant and a_n is the n th regression coefficient. The weighted values of each influencing factor were obtained by the product of y_n (characteristic value) and a_n (coefficient) of each influencing factor. And the landslide susceptibility index (IS) can be calculated as

$$IS = \frac{\exp(\mathrm{logit}(z))}{1 + \exp(\mathrm{logit}(z))} \tag{9}$$

3.3.4. Random forest (RF)

The RF model is an integrated learning method that combines multiple decision trees for classification and prediction. The classifier is a recursive process from the root node to the child nodes, selecting a random portion of samples and features from the training data with put-backs (Ilia et al., 2018; Dou et al., 2019). From the nodes of the tree, branches are determined based on the optimal features between the nodes, and branching is continued until the result of a tree is obtained. Finally, the subset of categories with the most votes can be selected as the final outputs:

$$T(X) = av_k \max_Y \sum_{i=1}^{k} I(t_i(X) = Y) \tag{10}$$

where $T(X)$ is the combined classification model, each node is denoted by k and t_i is the decision tree. The output variable and the feature function are represented by U and I. The marginal function can be denoted by the following equation:

$$mg(X,Y) = av_k I(t_k(X) = Y) - \max_{j \neq Y} av_k I(t_k(Y) = j) \tag{11}$$

The classification reliability of the model is proportional to the value of the function. The following equation is principle of the categorization:

$$P_{x,y}(P_\Theta(\mathrm{t}(X,\Theta) = Y) - \max P_\theta(h(X,\Theta) = j)) < 0 \tag{12}$$

where (X,U) is the probability space and P represents the feature variable.

The model does not require any transformation or rearrangement for disaggregated data, thus can effectively eliminate overfitting of data Considering it has strong generalization ability and high accuracy, the RF model has been widely used in the topic of LSA (Kim et al., 2018; Hong et al., 2019). The RF generally consists of two trees (positive and negative), each of which was constructed by using ten random features (influencing factors) in this study. The

purpose of this study is to explore the impact of sampling strategies on LSA, so using a stable model can reduce the effect of uncertainty from the model itself. This is the reason why the RF model was used as the approach to generate final landslide susceptibility maps.

3.4. Sampling strategy

3.4.1. Landslide sampling strategy

The conversion of landslide inventory from vector to raster is a delicate operation when conducting landslide susceptibility modelling. Different rasterization methods impact the calculation of landslide area which may result in large differences in landslide information. Moreover, the landslide shape, area and raster resolution also influence the final results, especially when the landslide area is small and the raster resolution is coarse (Arnone et al., 2016; Huang et al., 2022). When the number of pixels involved in the landslide boundary is large, whether to consider the landslide boundary pixels as the landslide sampling dataset may have a large impact on the landslide information included into the model. Hence, two landslide rasterization scenarios (landslide core and landslide extension) were designed to evaluate the influence of rasterization methods. Usually, the landslide boundary occupies one pixel, and sometimes up to two pixels (Fig. 4a). The landslide core was expressed by all the pixels contained in the landslide boundary and was obtained through the Polygon to Raster tool in GIS 10.6 software as shown in Fig. 4b. The landslide extension was the pixels combination of landslide core and landslide boundary (Fig. 4c). In this study, the landslide boundary includes 4161 pixels, the landslide core contains 9805 pixels, and the landslide extension contains 13966 pixels. The ratio of landslide boundary to core is 0.424 under the raster resolution of 30 m.

3.4.2. Non-landslide sampling strategy

Different from landslide samples, the sampling of non-landslide dataset is highly subjective. In literature, non-landslide samples were mostly randomly sampled from landslide-free area (Fig. 5a), thus this strategy was selected as the reference scenario in the present study. Regarding the controlling test, non-landslides were sampled from the buffer zone and the very low (VL) susceptibility zone generated from some models. The buffer zone is obtained by taking the centre of any landslide point and drawing a circle surface with the buffer distance as the radius, then taking the intersection of the circle surface. The buffer zone scenarios included the buffer distance to landslides of <200 m (Figs. 5b), 200–400 m (Fig. 5c) and outside of 400 m (Fig. 5d), which were designed to reveal the effect of buffer distances on the sampling.

Given that the LSA results mark different from model to model, two types of data-driven models, including generalized linear model (LR) and nonlinear regression model (C5.0 and SVM) were selected to generate landslide susceptibility maps in advance (pre-LSM) (Fig. 5e, f, and g). Then the VL zones in these pre-LSMs were used to prepare for formal non-landslide datasets. It should be noted that the pre-LSMs were generated by using the non-landslide samples from landslide-free area and combined with equal landslide samples.

Finally, fourteen scenarios were determined by combining the landslide and non-landslide sampling methods mentioned above, and the specific information are shown in Table 2. An equal number of landslide and non-landslide samples were selected to ensure a balance of positive and negative samples in the modelling dataset in SPSS Statistics software. Then, the LSIs of the recorded landslide samples were set to 1, whereas those of the randomly selected non-landslide samples were set to 0. As Shirzadi et al. (2017) suggested, the ratio of training to test datasets at 8:2 could obtain higher prediction accuracy in the raster resolution of 30 m. Thus, 80% of the modelling datasets (composed of equal landslide and non-landslide samples) under each scenario were randomly selected for the RF model training. The remaining 20% of the datasets were used for the test in SPSS modeler software.

3.5. Contribution of influencing factors

The evaluation of the importance of the influencing factors can reflect the reasonableness of the links established between the model and the actual geological environment, and help to establish a system of influencing factors suitable for landslide susceptibility analysis (Liang et al., 2021). In this study, the importance of ten influencing factors was assigned by RF model in SPSS modeler 18. Then each of the importance indices was normalized using the max–min method, which was calculated as the following equation:

$$f_i^* = \frac{f_i - f_{\min}}{f_{\max} - f_{\min}} \tag{13}$$

where f_i and f_i * denote the values before and after normalization respectively, $f_{\min}$ and $f_{\max}$ are the minimum and maximum values in the data.

3.6. Model performance evaluation

ROC curve is known to be used to evaluate the accuracy of binary classification tasks in machine learning models and is widely used in the analysis of LSA results (Rodrigues et al., 2021). The horizontal

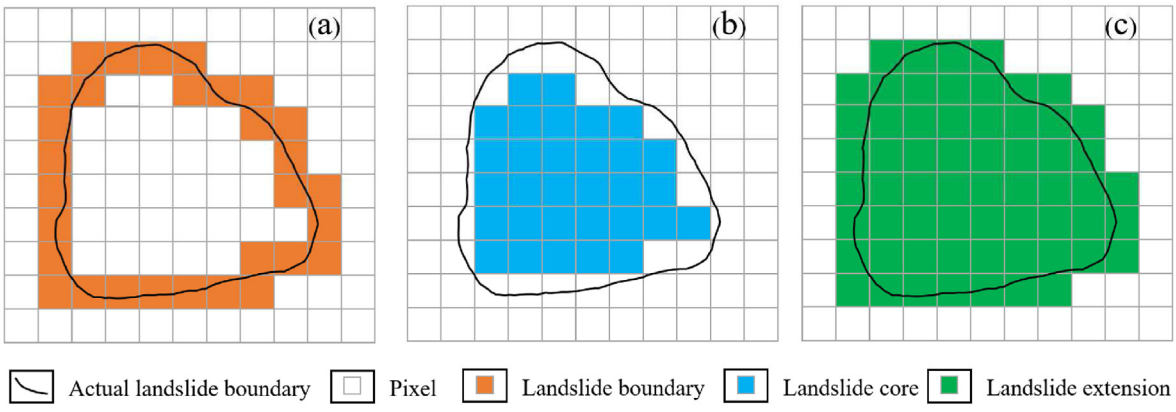

Fig. 4. Different rasterization ways to represent a landslide with pixels: (a) Landslide boundary, (b) Landslide core, and (c) Landslide extension.

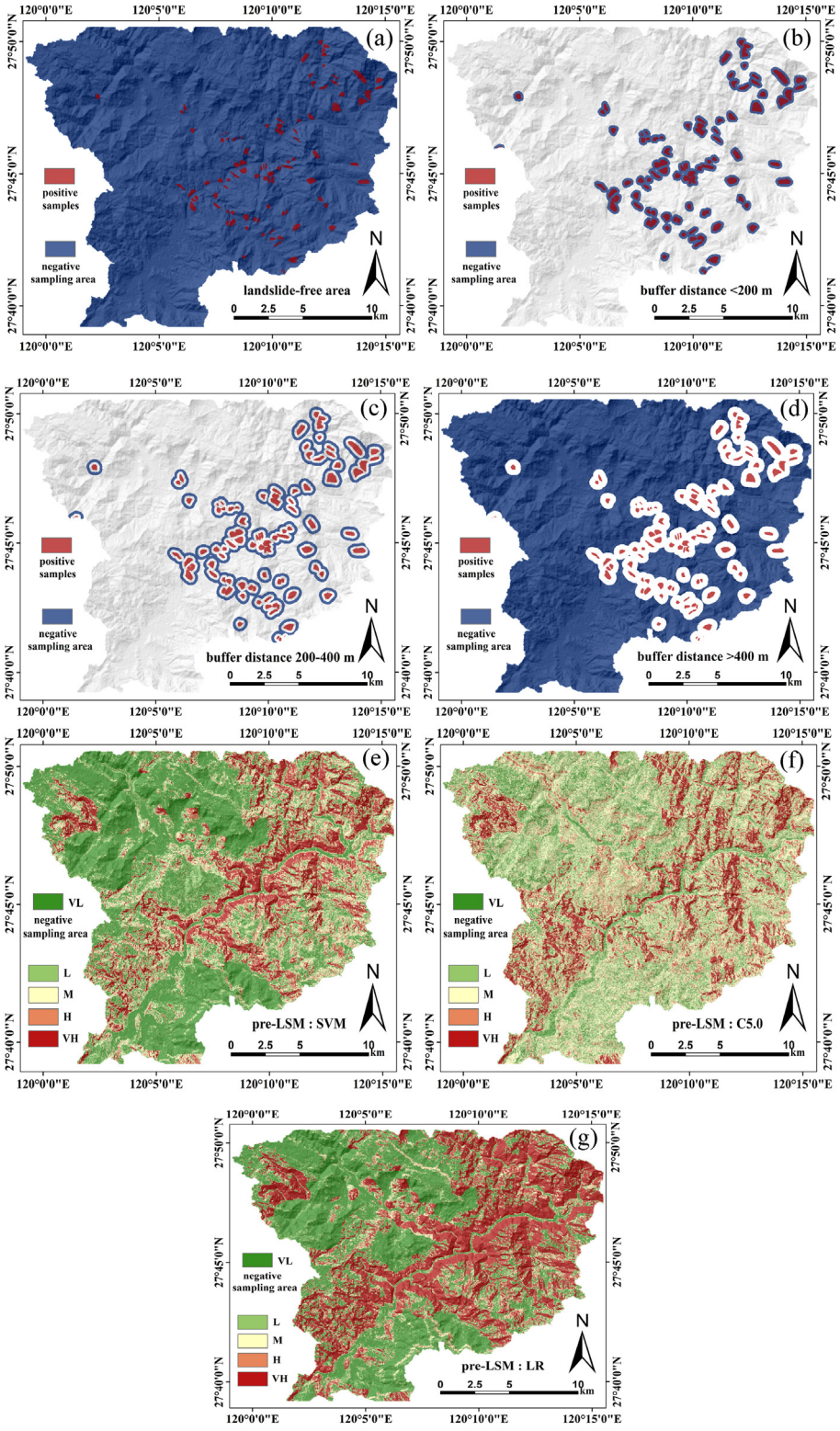

Fig. 5. The negative sampling area for each sampling strategy: (a) Landslide-free area, (b) Buffer distance <200 m, (c) Buffer distance 200–400 m, (d) Buffer distance >400 m; (e), (f) and (g) are the VL zone of pre-LSMs generated by SVM, C5.0 and LR models, respectively.

axis and vertical axis of the ROC curve are combined by 1-specificity and sensitivity, and the AUC of the ROC curve ranges from 0 to 1, closer to 1 indicates better model performance.

$$Sensitivity = \frac{TP}{TP + FN} \tag{14}$$

Table 2
The landslide and non-landslide sampling strategy of each scenario.

Landslide sampling	Non-landslide sampling (NLS)	Pixels of NLS selection area	Pixels of NLS	Percentage of NLS (%)
landslide core	VL zone from SVM	241,129	9805	0.41
	VL zone from C5.0	189,757	9805	5.17
	VL zone from LR	146,174	9805	6.71
	landslide-free area	421,503	9805	2.33
	buffer distance <200 m	32,346	9805	30.31
	buffer distance 200–400 m	36,449	9805	26.90
	buffer distance >400 m	350,977	9805	2.79
landslide extension	VL zone from SVM	241,129	13,966	0.58
	VL zone from C5.0	189,757	13,966	7.36
	VL zone from LR	146,174	13,966	9.55
	landslide-free area	417,342	13,966	3.35
	buffer distance <200 m	33,569	13,966	41.60
	buffer distance 200–400 m	36,531	13,966	38.23
	buffer distance >400 m	345,514	13,966	4.04

$$Specificity = \frac{FP}{TN + FP} \tag{15}$$

$$AUC = \frac{\sum_{i \in \text{positiveclass}} rank_i - \frac{M(1+M)}{2}}{M \times N} \tag{16}$$

where the value of $rank_i$ is the number of the i-th sample, M and N are the number of positive and negative samples, respectively.

4. Results

4.1. Factor importance and LSM

Based on the principles introduced before, the influencing factors were reclassified and the FR values of each category are shown in Table 3. We can see that the impacts of different factors and categories on the occurrence of landslides have evident differences. For instance, the FR values were all greater than 1 when the slope degrees larger than 37°, and increased with the increase of the degree. However, the FR values of the slope degrees lower than 37° were smaller than 1, thus indicating the negative effect on landslide occurrences in the study area. Regarding the soil factor, the FR values of all soil types were lower than 1 except the red soil, which indicated that this type of soil was the most important for the development of shallow landslides. This can be explained by the large porosity of red soil, thus its strength decreases more rapidly when exposed to water (Su et al., 2015). The classifications in land cover factor that had larger impacts on landslides included dryland (FR = 1.93), shrubs (FR = 1.09), open grass (FR = 1.50) and urban land (FR = 1.17). Regarding the distance to river, the overall trend is that the FR values decreased with the increase of distance rivers, which agreed well with the spatial distribution of landslides in the area.

As seen in Fig. 6, the importance of each influencing factor, which was represented by the importance measure (IM) value, was obtained from the RF model. We can see that six factors contributed more to the landslides in the study area, namely land cover (IM = 1), soil type (IM = 0.949), slope (IM = 0.923), distance to river (IM = 0.846), NDVI (IM = 0.769), elevation (IM = 0.641). The other four factors had relatively lower importance and their IM values were below 0.5. In general, the current results showed that the development of shallow landslides in the region was mainly controlled by the material (soil type), shape (slope), and surface covers (land cover, NDVI) of slopes. Moreover, no importance of the factors was smaller than 0, thus indicating it is reasonable to include these factors into the LSA of the study area.

Fourteen final LSMs were generated for different scenarios based on the RF model. The whole area was divided into five zones based on the natural breaks method, namely very-low susceptibility (VL), low susceptibility (L), moderate susceptibility (M), high susceptibility (H), and very high susceptibility (VH), as shown in Figs. 7 and 8. On the whole, although the LSMs under different scenarios are different, the distribution of VH, VL zones has similar patterns. The VH zones are mainly distributed at the river banks, since most settlements are located in this region, and human engineering activities can easily change slope stability. Another obvious VH zone is located in the mountainous area of the northeast which is characterized as higher altitude and larger relief. The

Table 3
The frequency ratio of each category within the influencing factors.

Factor	Category	FR	Factor	Category	FR	Factor	Category	FR
Aspect (°)	0–52	0.72	Relief	−34.0–4.1	0.13	NDVI	0.43–0.46	1.07
	52–126	1.15		−4.1–2.0	0.45		0.46–0.49	1.42
	126–199	1.41		−2.0–0.4	0.86		0.49–0.84	1.64
	199–251	1.22		0.4–2.1	1.29	Land cover	(1) Water Field	1.00
	251–306	0.81		2.1–3.3	1.52		(2) Dryland	1.93
	306–360	0.50		3.3–41	1.92		(3) Dense Forest	0.95
Elevation (m)	12–52	0.72	Slope (°)	0–13	0.14		(4) Shrubs	1.09
	52–180	1.90		13–21	0.37		(5) Open forest	0.93
	180–452	0.90		21–37	0.79		(6) Garden	0.87
	452–538	0.50		37–51	1.23		(7) Dense grass	0.65
	538–627	0.67		51–62	1.52		(8) Open grass	1.50
	627–1091	0.49		62–81	1.89		(9) Urban land	1.17
Profile Curvature	0.0–4.0	0.74	Distance to river (m)	100	1.69		(10) Water Area	0.00
	4.0–6.9	0.89		200	3.41		(11) Bare land	0.00
	6.9–9.8	1.02		300	3.15	Soil type	(1) red soil	1.45
	9.8–15.4	1.15		400	2.51		(2) yellow soil	0.16
	15.4–42.7	1.30		500	2.35		(3) paddy soil	0.86
TPI	−16.3–4.1	1.16		600	1.07		(4) submergic paddy soil	0.01
	−4.1–0.8	0.84		>600	0.62		(5) acid soil	0.55
	−0.8–1.0	0.88	NDVI	−0.30–0.31	0.27		(6) brown soil	0.10
	1.0–6.3	1.16		0.31–0.39	0.55			
	6.3–17.4	1.59		0.39–0.43	0.86			

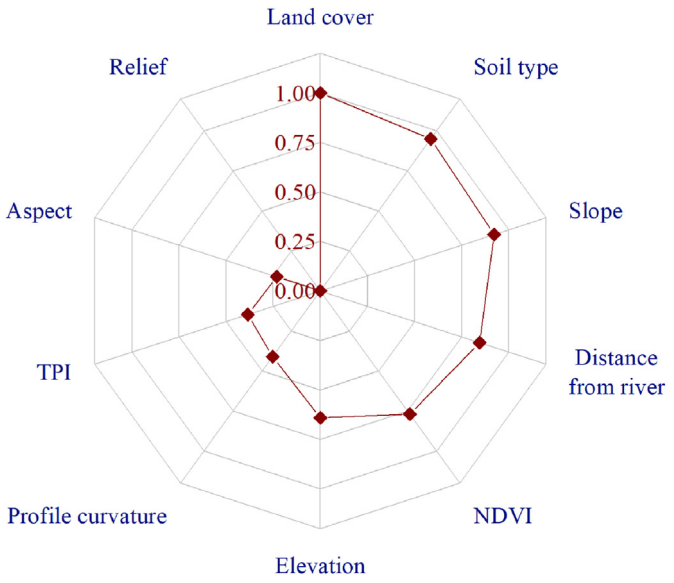

Fig. 6. Radar chart of the importance of the influencing factors, in which the numbers show the IM value of each factor obtained from the RF model.

VL area is mainly located in the central and northwestern part with relatively gentle terrain. Although the susceptibility zonation varies with scenarios, all these maps can generally reflect the development trend of landslides in the region. Regarding the comparison among different sampling scenarios, the maps using landslide core (Fig. 7) have more VH zone and less M zone than that using landslide extension dataset (Fig. 8). Moreover, when the buffer distance applied for non-landslide dataset increases, the VL and VH zones become larger (Fig. 7b–d, Fig. 8b–d).

4.2. Model performance evaluation and comparison

In order to reveal the distribution pattern of the susceptibility zonation and landslides in these LSMs, the area of each susceptibility level was first counted. The results show that the area of VH zone accounts for 12%–17% of the total area, and the area with medium susceptibility level accounts for 8%–30%. The area of L and VL zones are commonly larger than that of other susceptibility levels (Fig. 9a and b). However, this is not the case when using the non-landslide dataset from the VL-LR model. In this scenario, the area of VH zone is rather large, which reaches 39% and 42%, respectively. Next, we compared the percentage of landslides in each susceptibility level. When the landslide core is utilized as positive samples (Fig. 9c), the percentage of landslides identified in VL areas are no more than 2%, while this number can reach up to 90% in VH areas. In contrast, the landslide located in VL zone is less than 10% and the percentages of landslide in VH zones are between 44% and 83% when considering landslide extension as landslide dataset (Fig. 9d). Then, the indicator of frequency ratio (FR) was calculated to represent the relatively density degree of landslides in a specific susceptibility zone which considered both landslide numbers and total area. Generally speaking, a reasonable landslide susceptibility map is characterized as higher FR values in high and very high susceptibility areas. In this study, the FR values for all scenarios increase with the increase of susceptibility level (Fig. 9e and f). The FR values are less than 1 in low and very low susceptibility areas while larger than 1 in high and very high areas, thus indicating the landslide susceptibility zonation accurately captures the spatial distribution of historical landslides. It should be mentioned that the results regarding the scenario of VL-LR model is still an exception, where the FR values in H zone is less than 1, and FR in VH zones is evidently smaller than that in the other scenarios. This indicates that the performance by using the VL area from LR model as non-landslide dataset is not satisfactory.

Finally, the AUC values of training and testing datasets under 14 scenarios were calculated in SPSS modeler software to compare the performance of different sampling strategies (Fig. 10). The AUC values range from 0.733 to 0.942, thus indicating that all the scenarios have good predictive results. Moreover, the results obtained by using the training dataset and test dataset have similar accuracies, which verifies the generalization ability of the applied methods. Among all the scenarios, the peak accuracy is from the result by using landslide core and VL-C5.0 model with the AUC value of 0.933 (training dataset) and 0.942 (testing dataset).

4.3. Impact of sampling strategies on LSA

4.3.1. Impact of landslides sampling strategies

By comparing the landslide susceptibility maps of different scenarios, it can be found that the total area of H and VH zones of the maps applying landslide core as positive samples are slightly larger than those using the landslide extension dataset. Meanwhile, less L and VL zones are identified in the former maps than the latter ones. Moreover, the AUC values of the maps using landslide core are larger than those using landslide extension samples. Specifically, the maps using landslide core samples are larger than those using landslide extension by 0.2%–2.7% in AUC accuracy. These interesting results indicate that landslide cores can better reflect landslide characteristics, and better distinguish between landslide and non-landslide samples. This is mainly because that the landslide extension covers landslide boundary areas, which is the overlap zone between landslide and non-landslide. Pixels in these areas will dilute the landslide features and reduce the quality of landslide samples. Hence, it is recommended that users employ only landslide core as positive samples and discard pixels covering landslide boundary during landslide susceptibility modelling.

4.3.2. Impact of non-landslides sampling strategies

Considering the comparison results mentioned in the section above, the following analysis regarding non-landslide samples focus on the scenarios based on landslide core. Non-landslide sampling from landslide-free area is the most commonly used non-landslide sampling strategy, so this scenario is used as a reference to measure the magnitudes of change of modelling accuracy user the other scenarios. When the non-landslide dataset is obtained from the buffer zone around landslide boundary, the AUC accuracies improve with the increase of buffer distance. However, compared with the reference (AUC = 0.844), only the scenario when the buffer distance larger than 400 *m* has better performance (AUC = 0.878). The scenarios with the buffer distance of 200 m (AUC = 0.751) and 200–400 m (AUC = 0.801) are evidently lower in accuracy. This makes us conclude that the quality of non-landslide samples can be improved only when sampling at the area with a sufficient distance to landslide boundary. This can be explained that the pixels near from the landslide boundary have some similar characteristics with positive samples (landslide pixels) while the characteristics of negative samples (non-landslide pixels) are weakened. Therefore, the non-landslides dataset for machine learning model training are not representative enough.

Next, we assess the impacts of scenarios which use the VL zone from pre-LSM as non-landslide samples. It can be found that this sampling strategy can improve the performance of LSA. The AUC values are 0.857 (VL-SVM), 0.901 (VL-C5.0) and 0.839 (VL-LR), respectively. It should be mentioned that the performance of the scenario applying the VL-LR model is not superior to that of the

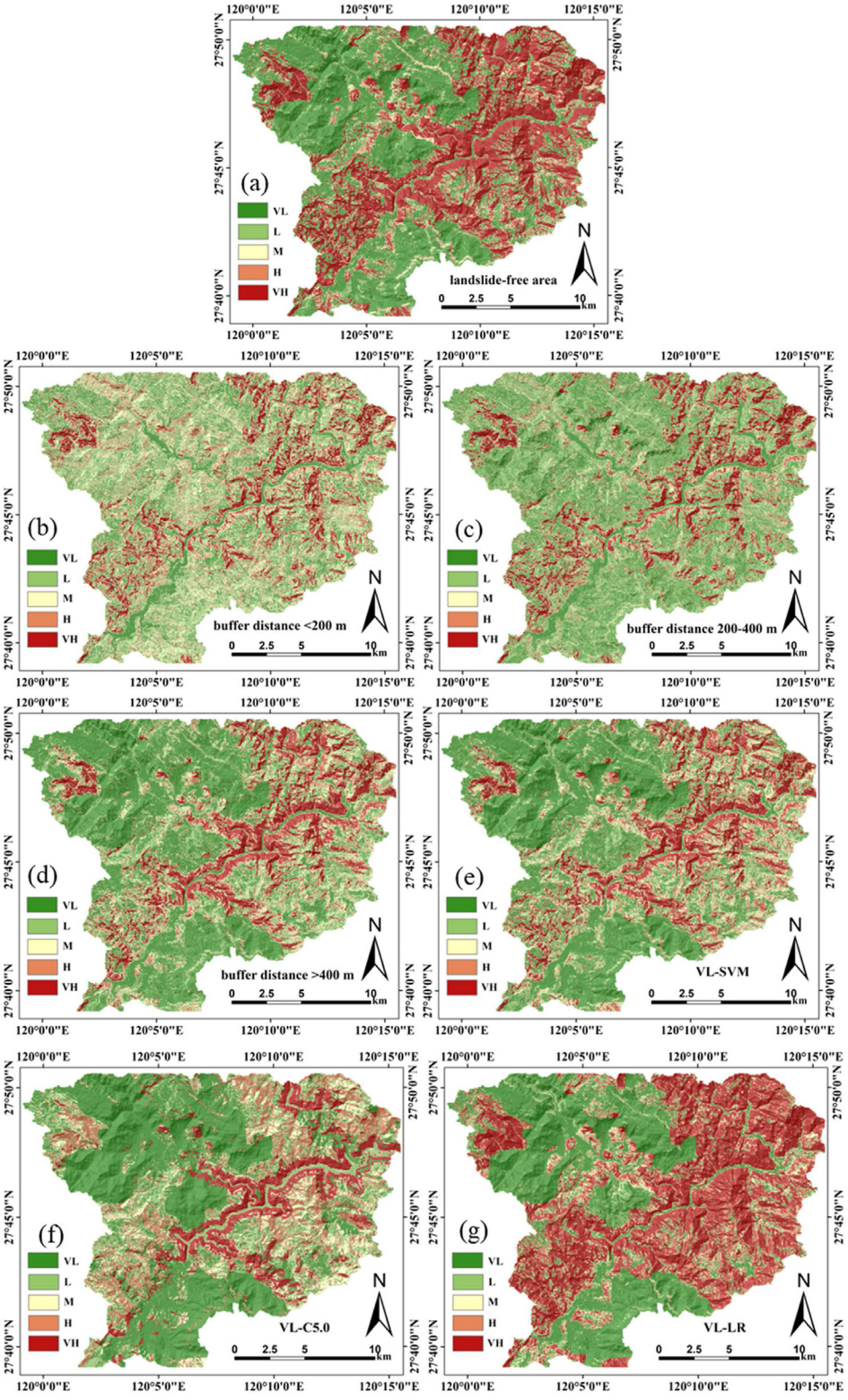

Fig. 7. The landslide susceptibility maps generated by using the landslide core as positive samples, where the negative dataset was sampled from: (a) Landslide-free area, (b) Buffer distance <200 m, (c) Buffer distance 200–400 m, (d) Buffer distance >400 m; (e), (f) and (g) are the VL zone of pre-LSMs obtained from the SVM, C5.0 DT and LR models, respectively.

reference, which supports the results from the statistics of landslide distribution. Under this scenario (Fig. 8g), the total area of high and very high susceptibility areas that were identified is larger than 50% of the entire region, which doesn't agree well with the actual situation. Similar results are not observed in the scenarios which determine the non-landslides through the VL-SVM and VL-C5.0 methods. From the perspective of model principles, the RF, SVM, and C5.0 models are nonlinear models, while the LR is a linear one.

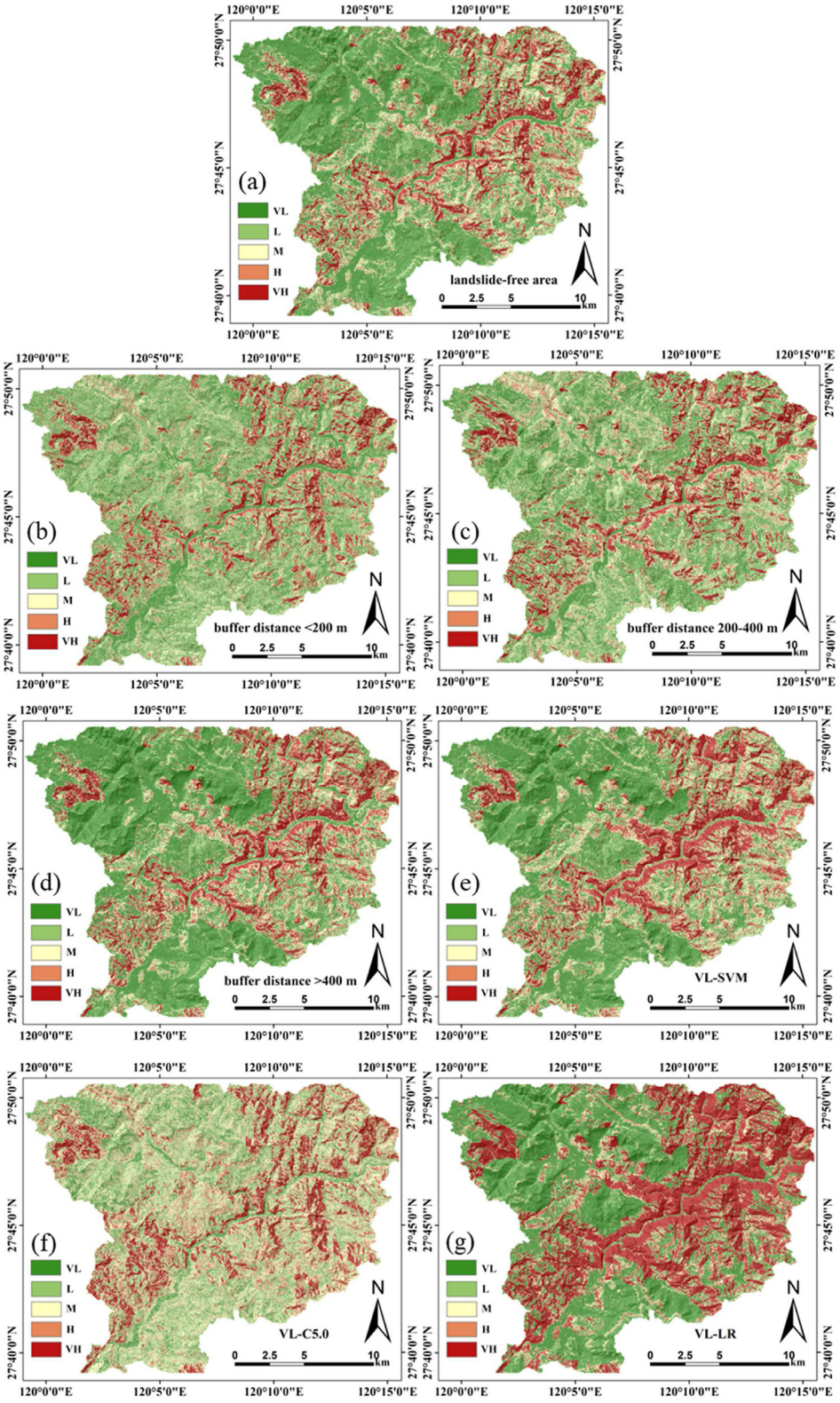

Fig. 8. The landslide susceptibility maps generated by using the landslide extension as positive samples, where the negative dataset was sampled from: (a) Landslide-free area, (b) Buffer distance <200 m, (c) Buffer distance 200– 400 m, (d) Buffer distance >400 m; (e), (f) and (g) are the VL zone of pre-LSMs obtained from the SVM, C5.0 DT and LR models, respectively.

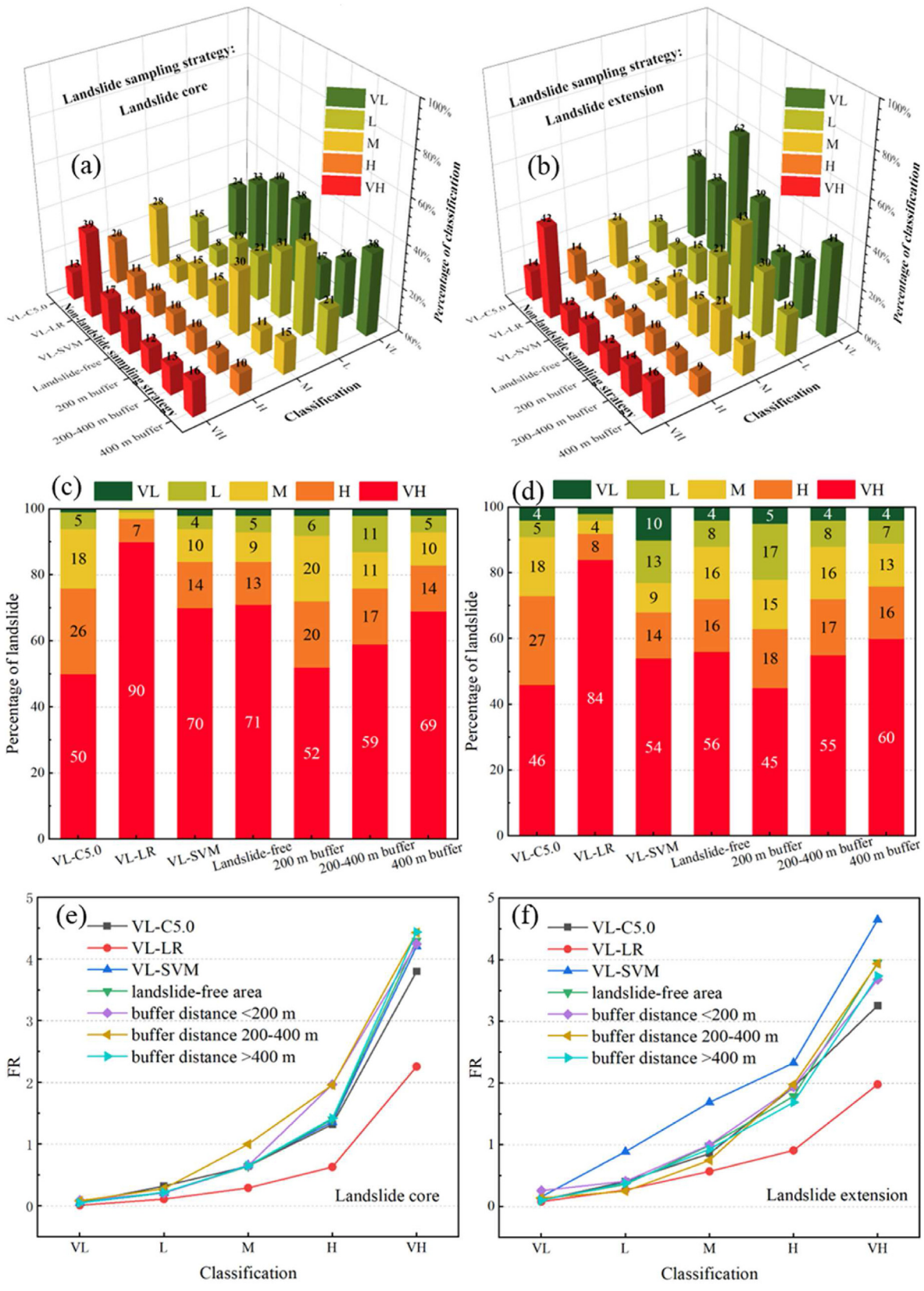

Fig. 9. The statistical indicators for the distribution of landslide susceptibility level and historical landslides: (a) The percentage of area of each landslide susceptibility level when using the landslide core dataset, (b) The percentage of area of each landslide susceptibility level when using the landslide extension dataset, (c) The percentage of landslides identified in each landslide susceptibility level when using the landslide core dataset, (d) The percentage of landslides identified in each landslide susceptibility level when using the landslide extension dataset, (e) The frequency ratio of historical landslides in each landslide susceptibility level when using the landslide core dataset, and (f) The frequency ratio of historical landslides in each landslide susceptibility level when using the landslide extension dataset.

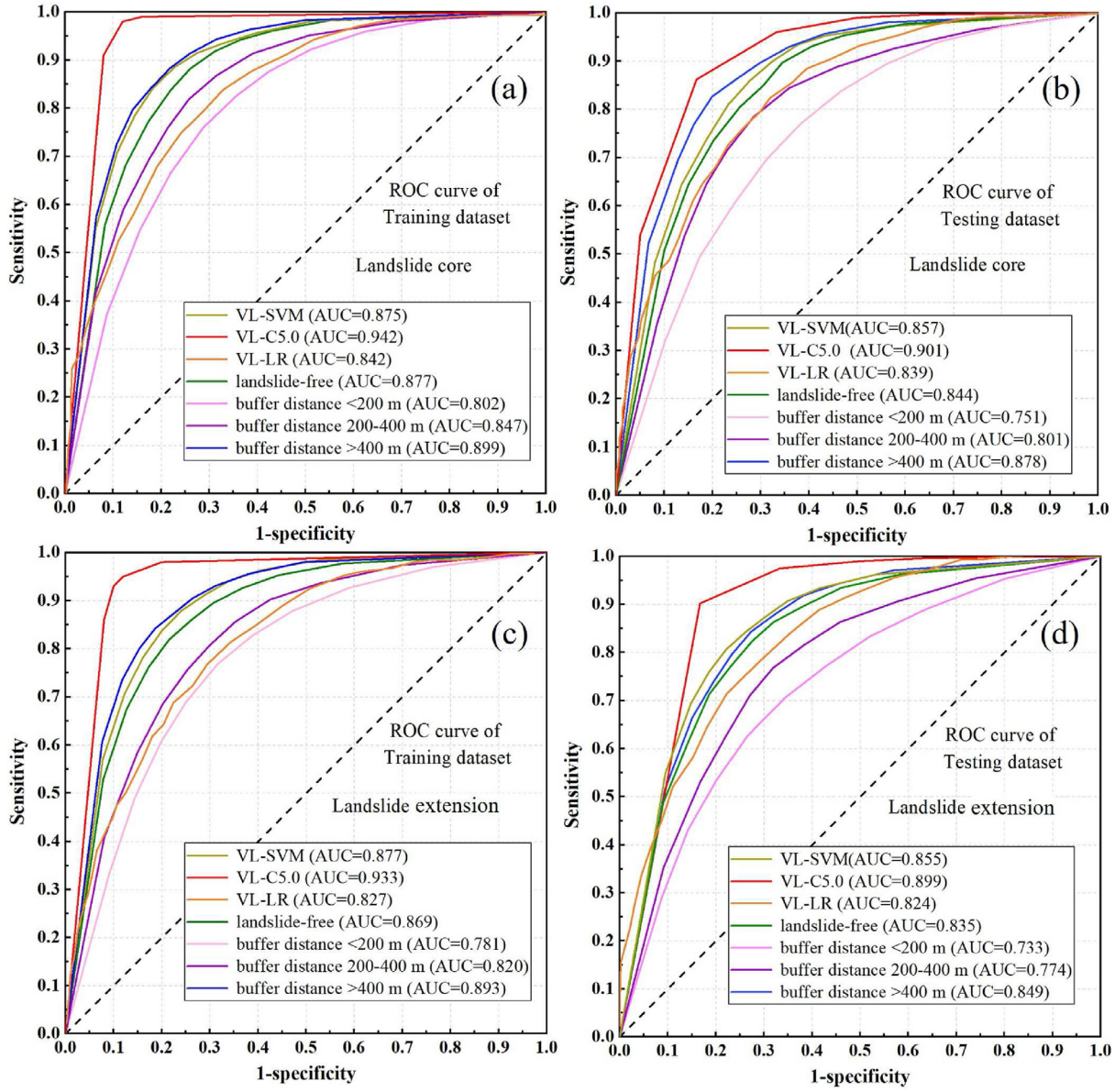

Fig. 10. The ROC curves and AUC values of different scenarios.: (a) Using landslide core and training dataset, (b) Using landslide core and testing dataset, (c) Using landslide extension and training dataset, and (d) Using landslide extension and testing dataset.

The former is commonly more complex than the latter since the classification criteria for linear classification are homogeneous. Hence, the scenarios using the VL-SVM and VL-C5.0 methods conduct two nonlinear classification (the second one is the final LSM by using the RF model), which cause higher model accuracies. From the perspective of sample quality, the very low susceptibility zones from the LR model only focus on a part of characteristics of non-landslide sample, which causes that the missing characteristics are identified as high and very high susceptibility zones in the LSM. This is also the reason why this scenario has larger VH and H susceptibility zones than the other scenarios. Overall, the negative dataset sampled from very low susceptibility zone determined by machine learning models can improve the performance of LSA compared with that from the landslide-free area, but it is recommended to select nonlinear models. Therefore, among all the scenarios in this study, the one applying landslide core (positive dataset) and VL zone from C5.0 model (negative dataset) has the highest accuracy.

5. Discussion

In this section three main points will be discussed: (i) the uncertainties associated with the modelling strategy and results, (ii) main limitations of the current method and findings, and (iii) the comparison with previous studies.

The uncertainties of this study are mainly related to one of the following aspects: (i) the relationship between data resolution and landslide size, (ii) determination of buffer distance between landslide area and landslide-free area, and (iii) selection of influencing factors for the LSA. The landslides in the study area are mostly small-scale in volume ($<10^5$ m^3), with the area ranging from 478 m^2 to 0.05 km^2. With the data resolution of 30 *m*, landslide boundary commonly accounts for a large proportion of the entire landslide, which can result in a significant difference in the total area of landslide datasets between the two positive sample strategies. The statistical results confirm this point: the number of landslide pixels obtained from the landslide core is higher than

those from the landslide extension by 40%. Hence, more fuzzy features related to the pixels covering landslide boundary are included into the positive dataset, leading to lower accuracies of the scenarios using landslide extension. It can be expected that the impact of sampling strategy of landslide dataset may decrease when it comes to large-scale landslides or higher resolution data. The choice of the proper resolution in LSM is always an operational issue, which is related to many aspects (Catani et al., 2013). Some studies test the performance of different data resolutions in regional LSM, but they failed to incorporate the variability of sampling strategy (e.g. Arnone et al., 2016; Schlögel et al., 2018). Hence, the future tests regarding the combined impacts of different scales, data resolution and sampling strategy may be of high interest.

Regarding the best definition of buffer distance for non-landslide sampling, there is no agreement on this topic so far. In this study, a measurement in GIS shows that the average distance among the landslides in the study area is approximately 400 *m*, thus we used the 200 *m* as the interval for buffer distance but set only three scenarios (<200 *m*, 200–400 *m*, >400 *m*). There are also studies using 1 hm (e.g. Xi et al., 2022) or kilometer (e.g. Gameiro et al., 2021) as interval and defining more than 5 different buffer distances. Evidently, more intervals can more accurately capture the impact of buffer distance, but our current outputs reveal similar findings with the ones previously mentioned, namely the accuracy of LSM becomes higher with the increase of distance between landslide and non-landslide samples. Moreover, it should be noted that the definition standard of buffer distance in this study is different from previous ones. We separate the landslide free area into several individual parts without interactive overlap, instead of setting the area with larger distance than a certain threshold as an entire (Fig. 11). Under this standard, we avoid potential repetitive sampling in landslide-free area.

The uncertainty originated from the influencing factors is dominantly associated with the model's predictivity ability. Many environmental factors have been incorporated in landslide susceptibility modelling, but a "perfect" combination of factors does not exist (van Westen et al., 2006). In this study, we selected 10 factors, which agree with previous studies for similar test sites (e.g. Su et al., 2015), and each factor has been verified as reasonable through the calculation of importance (Fig. 6). The results indicate that land use, slope, distance to river and NDVI are of higher importance for the landslide occurrence, thus partly supporting the point of Wang et al. (2020). However, there are studies revealing different findings where the important effect of rainfall on landslide susceptibility was emphasized (Su et al., 2015). In this study, the rainfall factor is discarded since it includes temporal information of landslides (Fell et al., 2008). Anyhow, the comparison regarding the factor importance makes us conclude that the contribution of factors may vary with local characteristics. Moreover, some studies show that eliminating or adding certain factors into a model may improve predictive ability (Pham et al., 2019; Tang et al., 2020), which was not conducted in this study because this is not our main objective.

Given the uncertainties mentioned above, the main limitations of the current method and findings can be summarized and stated. First, the influence of gridding methods of landslide samples on the model performance is greatly subject to the landslide area and type. However, limited by the type of slope failures in the study area, the current findings are only associated with shallow landslides. Second, the division of the buffer zone for the non-landslide samples is not detailed enough. Although the general law of the landslide susceptibility accuracy with the buffer distance was revealed, the optimal buffer distance has not been obtained. Finally, the results are not confirmed in other areas with different geological environments yet. Given that the accuracy of LSA is affected by many aspects, it is really difficult to directly predict the performance of the algorithm/model when it is applied for another region. We can only state that the performance of LSA is expected to be good when the proposed strategy is used for another similar area, when other procedures are normal. Therefore, it is necessary to verify the robustness and reliability of the obtained rules in other regions in future works. In spite of all these drawbacks, our results show that the uncertainty during the modelling process is acceptable: The AUC accuracy of all the scenarios is higher than 0.73 by using the testing dataset, and most scenarios have an accuracy larger than 0.8. This allows us to focus on the analysis of the impacts of sampling strategy instead of the improvement of model performance.

This study conducts an analysis of the combined impacts of landslide/non-landslide sampling strategy. It is difficult to directly

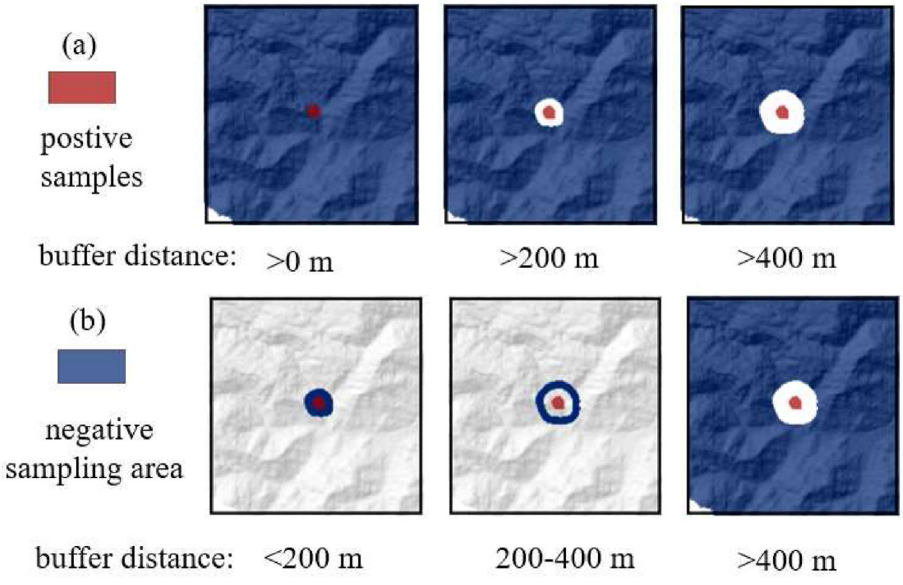

Fig. 11. The comparison of buffer zone around landslide pixels: (a) Traditional buffer area used in the literature, and (b) The scenario used in this study.

compare our work to the other studies, because they often only considered one separate sampling strategy. For example, Dou et al. (2020) compared the performance of four different sampling techniques in LSM, and found that the order of predictive power is landslide scarp > landslide body > centroid of scarp > centroid of the body. Huang et al. (2022) tested the effects of different spatial shapes of landslide boundary, and confirmed better performance of landslide polygon in LSA. However, the investigation regarding the landslide expression was missing in the literature described above. Although the objectives are different, it seems that their works are useful supplementary for us, and a completed framework on landslide sampling strategy for different scenarios can be generated by combining current findings. Regarding the non-landslide sampling strategy, Lucchese et al. (2021) tested the performance by obtaining non-landslide samples from buffer zone and known lowlands, and found that a priori intervention on non-landslide samples (i.e., sample from lowlands) can produce higher accuracy but are improper for generalization. This procedure needs rich expert knowledge to determine known non-susceptible area, but this is not available in the Feiyun catchment. Xi et al. (2022) compared the influences of the traditional buffer-controlled sampling method and a Newmark-based sampling approach for earthquake-triggered landslides, and found the latter one generated better results. Some other studies also have made attempts to obtain negative samples by machine learning models, including fractal theory (Hu et al., 2020) and self-organizing neural networks (Huang et al., 2017). Interestingly, their results all reveal that new strategies for sampling non-landslides can obtain higher predictive ability than traditional methods, which supports our results. Hence, it is necessary to develop and test more non-landslide sampling techniques as alternatives of traditional ones in the future, which would certainly provide new insights on this topic.

6. Conclusions

Regional LSA and associated uncertainties are one of the major challenges for landslide risk management and reduction. In the present study, the Feiyun catchment in southeast China was selected as a study area to test the impacts of sampling strategies of landslide and non-landslide datasets on the performance of LSM. Our results indicate that when the pixels covering only the landslide core are used as positive samples, the accuracy of LSM is higher than that of the map applying the pixels covering both landslide core and landslide boundary, with the improvement magnitude from 0.2% to 2.7%. A comparison regarding the non-landslide sampling strategies showed that the negative samples from the very low susceptibility identified by nonlinear machine learning models can also improve the susceptibility modelling performance. Hence, the commonly used strategy which selects non-landslide samples from landslide-free areas may enlarge the uncertainty for the modelling. The results have demonstrated also that, the LSA can also have better performance when the non-landslide dataset is determined from the buffer zone around landslide pixels. However, it should be noted that this improvement is closely related to the selection of buffer distance: when the buffer distance is less than 400 m, the accuracy of the landslide susceptibility map decreases instead. By combining different positive and negative sampling strategies, we determined 14 scenarios to generate regional landslide susceptibility maps. Among all the scenarios, the best predictive capability has been attributed to the landslide core (positive sample) and the very low susceptible zone from the C5.0 model (negative sample), with a peak AUC accuracy of 0.901.

Overall, the current test confirms that some uncertainties in LSA are associated with the dataset sampling strategy, but they can be reduced by improving the quality of positive and negative datasets. Hence, it is recommended to incorporate reasonable strategies regarding dataset sampling into the framework of LSM to make the outputs more reliable. Last, our future works can focus on the role of sampling strategy under different landslide sizes and data resolution.

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.

Acknowledgments

This research is funded by National Natural Science Foundation of China (Grant No 42307248) and Natural Science Foundation of Hebei Province (Grant No. D2022202005). Bixia Tian wants to thank the support from the Graduated Student Innovation Funding Project of Hebei Province (Grant No. CXZZSS2024007).

References

Agterberg, F., 2022. How can Earth science help reduce the adverse effects of climate change? J. Earth Sci. 33 (5), 1338-1338.
Ali, S., Parvin, F., Pham, Q.B., Khedher, K.M., Dehbozorgi, M., Rabby, Y.W., Anh, D.T., Nguyen, D.H., 2022. An ensemble random forest tree with SVM, ANN, NBT, and LMT for landslide susceptibility mapping in the Rangit River watershed, India. Nat. Hazards 113, 1601–1633.
Arnone, E., Francipane, A., Scarbaci, A., Puglisi, C., Noto, L.V., 2016. Effect of raster resolution and polygon-conversion algorithm on landslide susceptibility mapping. Environ. Model. Software 84, 467–481.
Azarafza, M., Ghazifard, A., Akgün, H., Asghari-Kaljahi, E., 2018. Landslide susceptibility assessment of south pars special zone, southwest Iran. Environ. Earth Sci. 77 (24), 1–29.
Azarafza, M., Azarafza, M., Akgün, H., Atkinson, P.M., Derakhshani, R., 2021. Deep learning-based landslide susceptibility mapping. Sci. Rep. 11 (1), 1–16.
Barik, M.G., Adam, J.C., Barber, M.E., Muhunthan, B., 2017. Improved landslide susceptibility prediction for sustainable forest management in an altered climate. Eng. Geol. 230, 104–117.
Brenning, A., 2005. Spatial prediction models for landslide hazards: review, comparison and evaluation. Nat. Hazards Earth Syst. Sci. 5, 853–862.
Bueechi, E., Klimeš, J., Frey, H., Huggel, C., Strozzi, T., Cochachin, A., 2019. Regional scale landslide susceptibility modelling in the Cordillera Blanca, Peru—a comparison of different approaches. Landslides 16, 395–407.
Bui, D.T., Pradhan, B., Lofman, O., Revhaug, I., 2012. Landslide susceptibility assessment Vietnam using support vector machines, decision tree, and Naïve Bayes models. Math. Probl Eng. 6, 1–26.
Bui, D.T., Shahabi, H., Omidvar, E., Shirzadi, A., Geertsema, M., Clague, J., Khosravi, K., Pradhan, B., Pham, B.T., Chapi, K., Barati, Z., Ahmad, B.B., Rahmani, H., Gróf, G., Lee, S., 2019. Shallow landslide prediction using a novel hybrid functional machine learning algorithm. Rem. Sens. 11 (8), 931.
Catani, F., Lagomarsino, D., Segoni, S., Tofani, V., 2013. Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. Nat. Hazards Earth Syst. Sci. 13 (11), 2815–2831.
Chen, W., Li, Y., 2020. Gis-based evaluation of landslide susceptibility using hybrid computational intelligence models. Catena 195, 104777.
Conforti, M., Pascale, S., Robustelli, G., Sdao, F., 2014. Evaluation of prediction capability of the artificial neural networks for mapping landslide susceptibility in the Turbolo River catchment (northern Calabria, Italy). Catena 113, 236–250.
Dao, D.V., Jaafari, A., Bayat, M., Mafi-Gholami, D., Qi, C.C., Moayedi, H., Phong, T.V., Ly, H.B., Le, T.T., Trinh, P.T., Luu, C., Quoc, N.K., Thanh, B.N., Pham, B.T., 2020. A spatially explicit deep learning neural network model for the prediction of landslide susceptibility. Catena 188, 104451.
Dou, J., Yunus, A.P., Bui, D.T., Merghadi, A., Sahana, M., Zhu, Z.F., Chen, C.W., Khosravi, K., Yang, Y., Pham, B.T., 2019. Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan. Sci. Total Environ. 662, 332–346.
Dou, J., Yunus, A.P., Merghadi, A., Shirzadi, A., Nguyen, H., Hussain, Y., Avtar, R., Chen, Y., Pham, B.T., Yamagishi, H., 2020. Different sampling strategies for predicting landslide susceptibilities are deemed less consequential with deep learning. Sci. Total Environ. 720, 137320.
Eker, A.M., Dikmen, M., Cambazoğlu, S., Düzgün, H.S.B., Akgün, H., 2015. Evaluation and comparison of landslide susceptibility mapping methods: a case study for the ulus district, bartın, northern Turkey. Int. J. Geogr. Inf. Sci. 29 (1), 132–158.
Fell, R., Corominas, J., Bonnard, C., Cascini, L., Leroi, E., Savage, W.Z., Jtc-Joint-Tech, C.L.E., 2008. Guidelines for landslide susceptibility, hazard and risk-zoning for land use planning. Eng. Geol. 102 (3–4), 85–98.

Galli, M., Ardizzone, F., Cardinali, M., Guzzetti, F., Reichenbach, P., 2008. Comparing landslide inventory maps. Geomorphology 94 (3–4), 268–289.
Gameiro, S., Riffel, E.S., de Oliveira, G.G., Guasselli, L.A., 2021. Artificial neural networks applied to landslide susceptibility: the effect of sampling areas on model capacity for generalization and extrapolation. Appl. Geogr. 137, 102598.
Goyes-Penafiel, P., Hernandez-Rojas, A., 2021. Landslide susceptibility index based on the integration of logistic regression and weights of evidence: a case study in Popayan, Colombia. Eng. Geol. 280, 105958.
Guo, Z., Chen, L., Yin, K., Shrestha, D.P., Zhang, L., 2020a. 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, 105667.
Guo, Z., Chen, L., Gui, L., Du, J., Yin, K., Do, H.M., 2020b. Landslide displacement prediction based on variational mode decomposition and WA-GWO-BP model. Landslides 17, 567–583.
Guo, Z., Shi, Y., Huang, F., Fan, X., Huang, J., 2021. Landslide susceptibility zonation method based on C5.0 decision tree and K-means cluster algorithms to improve the efficiency of risk management. Geosci. Front. 12 (6), 101249.
Guo, Z., Torra, O., Hürlimann, M., Medina, V., Puig-Polo, C., 2022. FSLAM: a QGIS plugin for fast regional susceptibility assessment of rainfall-induced landslides. Environ. Model. Software 150, 105354.
Guo, Z., Tian, B., He, J., Xu, C., Zeng, T., Zhu, Y., 2023a. Hazard assessment for regional typhoon-triggered landslides by using physically-based model -A case study from southeastern China. Georisk 17 (4), 740–754.
Guo, Z., Tian, B., Li, G., Huang, D., Zeng, T., He, J., Song, D., 2023b. 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, 1033085.
Guzzetti, F., Mondini, A.C., Cardinali, M., Fiorucci, F., Santangelo, M., Chang, K.T., 2012. Landslide inventory maps: new tools for an old problem. Earth Sci. Rev. 112 (1–2), 42–66.
He, Q., Shahabi, H., Shirzadi, A., Li, S., Chen, W., Wang, N., Chai, H., Bian, H., Ma, J., Chen, Y., Wang, X., Chapi, K., Ahmad, B.B., 2019. Landslide spatial modelling using novel bivariate statistical based Naïve Bayes, RBF Classifier, and RBF Network machine learning algorithms. Sci. Total Environ. 663, 1–15.
Hong, H.Y., Miao, Y.M., Liu, J.Z., Zhu, A.X., 2019. Exploring the effects of the design and quantity of absence data on the performance of random forest-based landslide susceptibility mapping. Catena 176, 45–64.
Hong, H.Y., Liu, J.Z., Zhu, A.X., 2020. Modeling landslide susceptibility using LogitBoost alternating decision trees and forest by penalizing attributes with the bagging ensemble. Sci. Total Environ. 718, 137231.
Hu, Q., Zhou, Y., Wang, S., Wang, F., 2020. Machine learning and fractal theory models for landslide susceptibility mapping: case study from the Jinsha River Basin. Geomorphology 251, 106975.
Huang, F.M., Yin, K.L., Huang, J.S., Gui, L., Wang, P., 2017. Landslide susceptibility mapping based on self-organizing-map network and extreme learning machine. Eng. Geol. 223, 11–22.
Huang, F.M., Ye, Z., Jiang, S.H., Huang, J., Chang, Z., Chen, J., 2020. Uncertainty study of landslide susceptibility prediction considering the different attribute interval numbers of environmental factors and different data-based models. Catena 202, 105250.
Huang, F.M., Yan, J., Fan, X.M., Yao, C., Huang, J.S., Chen, W., Hong, H.Y., 2022. Uncertainty pattern in landslide susceptibility prediction modelling: effects of different landslide boundaries and spatial shape expressions. Geosci. Front. 13 (2), 101317.
Hungr, O., Leroueil, S., Picarelli, L., 2014. The Varnes classification of landslide types, an update. Landslides 11 (2), 167–194.
Hürlimann, M., Guo, Z.Z., Puig-Polo, C., Medina, V., 2022. Impacts of future climate and land cover changes on landslide susceptibility: regional scale modelling in the Val d'Aran region (Pyrenees, Spain). Landslides 19 (1), 99–118.
Ilia, I., Loupasakis, C., Tsangaratos, P., 2018. Land subsidence phenomena investigated by spatiotemporal analysis of groundwater resources, remote sensing techniques, and random forest method: the case of Western Thessaly, Greece. Environ. Monit. Assess. 190 (11), 623.
Kavzoglu, T., Sahin, E.K., Colkesen, I., 2014. Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. Landslides 11 (3), 425–439.
Kim, J.C., Lee, S., Jung, H.S., Lee, S., 2018. Landslide susceptibility mapping using random forest and boosted tree models in Pyeong-Chang, Korea. Geocarto Int. 33 (9), 1000–1015.
Liang, Z., Wang, C.M., Khan, K., 2021. Application and comparison of different ensemble learning machines combining with a novel sampling strategy for shallow landslide susceptibility mapping. Stoch. Environ. Res. Risk Assess. 35 (6), 1243–1256.
Liu, F., Zhang, G.L., Song, X., Li, D., Zhao, Y., Yang, J., Wu, H., Yang, F., 2020. High-resolution and three-dimensional mapping of soil texture of China. Geoderma 361, 114061.
Lucchese, L.V., de Oliveira, G.G., Pedrollo, O.C., 2021. Investigation of the influence of nonoccurrence sampling on landslide susceptibility assessment using Artificial Neural Networks. Catena 198, 105067.
Ma, P., Peng, J., Zhuang, J., Zhu, X., Liu, C., Cheng, Y., Zhang, Z., 2022a. Initiation mechanism of loess mudflows by flume experiments. J. Earth Sci. 33 (5), 1166–1178.
Ma, S., Shao, X., Xu, C., 2022b. Characterizing the distribution pattern and a physically based susceptibility assessment of shallow landslides triggered by the 2019 heavy rainfall event in longchuan county, guangdong Province, China. Rem. Sens. 14 (17), 4257.
Medina, V., Hürlimann, M., Guo, Z., Lloret, A., Vaunat, J., 2021. Fast physically-based model for rainfall-induced landslide susceptibility assessment at regional scale. Catena 201, 105213.
Mehrabi, M., 2022. Landslide susceptibility zonation using statistical and machine learning approaches in Northern Lecco. Italy. Nat. Hazards 111 (1), 901–937.
Merghadi, A., Yunus, A.P., Dou, J., Whiteley, J., ThaiPham, B., Bui, D.T., Avtar, R., Abderrahmane, B., 2020. Machine learning methods for landslide susceptibility studies: a comparative overview of algorithm performance. Earth Sci. Rev. 207, 103225.
Moosavi, V., Talebi, A., Shirmohammadi, B., 2014. Producing a landslide inventory map using pixel-based and object-oriented approaches optimized by Taguchi method. Geomorphology 204, 646–656.
Nikoobakht, S., Azarafza, M., Akgün, H., Derakhshani, R., 2022. Landslide susceptibility assessment by using convolutional neural network. Appl. Sci. 12 (12), 5992.
Okalp, K., Akgün, H., 2016. National level landslide susceptibility assessment of Turkey utilizing public domain dataset. Environ. Earth Sci. 75 (9), 847.
Okalp, K., Akgün, H., 2022. Landslide susceptibility assessment in medium-scale: case studies from the major drainage basins of Turkey. Environ. Earth Sci. 81 (8), 244.
Paryani, S., Neshat, A., Pradhan, B., 2021. Improvement of landslide spatial modeling using machine learning methods and two Harris hawks and bat algorithms. Egypt. J. Remote Sens. Space Sci. 24 (3), 845–855.
Peng, L., Niu, R.Q., Huang, B., Wu, X.L., Zhao, Y.N., Ye, R.Q., 2014. Landslide susceptibility mapping based on rough set theory and support vector machines: a case of the Three Gorges area, China. Geomorphology 204, 287–301.
Petley, D.N., 2012. Global patterns of loss of life from landslides. Geology 40, 927–930.
Pham, B.T., Pradhan, B., Bui, D.T., Prakash, I., Dholakia, M.B., 2016. A comparative study of different machine learning methods for landslide susceptibility assessment: a case study of Uttarakhand area (India). Environ. Model. Software 84, 240–250.
Pham, B.T., Jaafari, A., Prakash, I., Bui, D.T., 2019. A novel hybrid intelligent model of support vector machines and the MultiBoost ensemble for landslide susceptibility modeling. Bull. Eng. Geol. Environ. 78, 2865–2886.
Pradhan, B., 2010. Landslide susceptibility mapping of a catchment area using frequency ratio, fuzzy logic and multivariate logistic regression approaches. J. Indian Soc. Remote Sens. 38 (2), 301–320.
Reichenbach, P., Rossi, M., Malamud, B.D., Mihir, M., Guzzetti, F., 2018. A review of statistically-based landslide susceptibility models. Earth Sci. Rev. 180, 60–91.
Rodrigues, S.G., Silva, M.M., Alencar, M.H., 2021. A proposal for an approach to mapping susceptibility to landslides using natural language processing and machine learning. Landslides 18 (7), 2515–2529.
Schlögel, R., Marchesini, I., Alvioli, M., Reichenbach, P., Rossi, M., Malet, J-P. Optimizing landslide susceptibility zonation: effects of DEM spatial resolution and slope unit delineation on logistic regression models. Geomorphology. 301, 10-20.
Sezer, E.A., Nefeslioglu, H.A., Osna, T., 2017. An expert-based landslide susceptibility mapping (LSM) module developed for Netcad Architect Software. Comput. Geosci. 98, 26–37.
Shahabi, H., Hashim, M., 2015. Landslide susceptibility mapping using GIS-based statistical models and Remote sensing data in tropical environment. Sci. Rep. 5, 9899.
Shahri, A.A., Spross, J., Johansson, F., Larsson, S., 2019. Landslide susceptibility hazard map in Southwest Sweden using artificial neural network. Catena 183, 104225.
Shirzadi, A., Chapi, K., Shahabi, H., Solaimani, K., Kavian, A., Ahmad, B.B., 2017. Rock fall susceptibility assessment along a mountainous road: an evaluation of bivariate statistic, analytical hierarchy process and frequency ratio. Environ. Earth Sci. 76, 152.
Smith, H.G., Spiekermann, R., Betts, H., Neverman, A.J., 2021. Comparing methods of landslide data acquisition and susceptibility modelling: examples from New Zealand. Geomorphology 3, 107660.
Su, C., Wang, L.L., Wang, X.Z., Huang, Z.C., Zhang, X.C., 2015. Mapping of rainfall-induced landslide susceptibility in Wencheng, China, using support vector machine. Nat. Hazards 76 (3), 1759–1779.
Su, A., Feng, M., Dong, S., Zou, Z., Wang, J., 2022. Improved statically solvable slice method for slope stability analysis. J. Earth Sci. 33 (5), 1190–1203.
Tang, Y., Feng, F., Guo, Z., Feng, W., Li, Z., Wang, J., Sun, Q., Ma, H., Li, Y., 2020. 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, 124159.
van Westen, C.J., van Asch, T.W.J., Soeters, R., 2006. Landslide hazard and risk zonation—why is it still so difficult? Bull. Eng. Geol. Environ. 65, 167–184.
van Westen, C.J.V., Castellanos, E., Kuriakose, S.L., 2008. Spatial data for landslide susceptibility, hazard, and vulnerability assessment: an overview. Eng. Geol. 102 (3–4), 112–131.
Varnes, D.J., 1978. Slope movement types and processes. In: Schuster, R.L., Krizek, R.J. (Eds.), Landslides: Analysis and Control, National Research Council, vol. 176. Transportation Research Board, National Academy Press, Special Report, Washington, D.C., pp. 11–33

Wang, Y.M., Feng, L.W., Li, S.J., Ren, F., Du, Q.Y., 2020. A hybrid model considering spatial heterogeneity for landslide susceptibility mapping in Zhejiang Province, China. Catena 188, 104425.
Xi, C.J., Han, M., Hu, X.W., Liu, B., He, K., Luo, G., Cao, X.C., 2022. Effectiveness of Newmark-based sampling strategy for coseismic landslide susceptibility mapping using deep learning, support vector machine, and logistic regression. Bull. Eng. Geol. Environ. 81 (5), 174.
Zêzere, J.L., Pereira, S., Melo, R., Oliveira, S.C., Garcia, R., 2017. Mapping landslide susceptibility using data-driven methods. Sci. Total Environ. 589, 250–267.
Zhang, T.-y., Han, L., Zhang, H., Zhao, Y.-H., Li, X.-a., Zhao, L., 2019. GIS-based landslide susceptibility mapping using hybrid integration approaches of fractal dimension with index of entropy and support vector machine. J. Mt. Sci. 16 (6), 1275–1288.
Zhao, Y., Wang, R., Jiang, Y., Liu, H., Wei, Z., 2019. GIS-based logistic regression for rainfall-induced landslide susceptibility mapping under different grid sizes in Yueqing, Southeastern China. Eng. Geol. 259, 105147.
Zhou, J., Cui, P., Hao, M., 2016. Comprehensive analyses of the initiation and entrainment processes of the 2000 Yigong catastrophic landslide in Tibet, China. Landslides 13, 39–54.
Zhu, A.X., Miao, Y., Liu, J., Bai, S., Zeng, C., Ma, T., Hong, H., 2019. A similarity-based approach to sampling absence data for landslide susceptibility mapping using data-driven methods. Catena 183, 104188.

Dr. Zizheng Guo is an associate professor of School of Civil and Transportation Engineering in Hebei University of Technology. He obtained his B Sc. and Ph.D. degrees in Geological Engineering from China University of Geosciences (Wuhan) in 2016 and 2021, respectively. His research interests include: (i) slope monitoring and stability evaluation, and (ii) landslide susceptibility and risk assessment.