Arabian Journal of Geosciences (2021) 14: 440
https://doi.org/10.1007/s12517-021-06630-5

ORIGINAL PAPER

A comparative study of the bivariate, multivariate and machine-learning-based statistical models for landslide susceptibility mapping in a seismic-prone region in China

Suhua Zhou[1,2] · Yunqiang Zhang[1] · Xin Tan[1,2] · Syed Muntazir Abbas[3]

Received: 12 February 2020 / Accepted: 23 January 2021 / Published online: 6 March 2021

Abstract
Statistical landslide susceptibility mapping (LSM) models have been most widely used in literatures. However, limitations and uncertainties remain in these methods. The main goal of the current study was to test and compare the efficiency of a bivariate model (the weight of evidence (WoE)), a multivariate model (logistic regression (LR)) and a machine-learning algorithm (the support vector machine (SVM)) in LSM. Lushan County of China was chosen because of its mountainous terrain and high risky of devastating seismic activities. An inventory of 867 landslides was utilized in this study, 70% of which were used to train these models, and the rest 30% were used to validate their accuracies. Ten factors of aspect, elevation, slope, curvature, peak ground acceleration (PGA), distance to the river (DtoR), lithology, topographic wetness index (TWI), stream power index (SPI) and percentage of tree cover (PTC) were used as input of the landslide susceptibility mapping (LSM) models. Accuracy evaluation based on the areas under the receiver operating characteristic curves (AUC) showed that the LR model gives the highest success rate (78.2%) and prediction rate (76.4%), the SVM has the second-highest success rate (75.9%) and the WoE had the second-highest prediction rate (75.6%). Comparison results suggested that the LR and the SVM are proper models for LSM of the study area. The obtained susceptibility maps would benefit regional land planning and seismic landslide hazard mitigation in the study area.

Keywords Landslide susceptibility mapping · GIS · Seismic-prone region · Weight of evidence · Logistic regression · Support vector machine

Introduction

Landslide represents a major geohazard type globally, inducing a great many casualties and immense property losses every year. Taking China as an example, an average of 6000 landslides occurred in China, causing an estimated property loss of 5–8 billion CNY and hundreds of deaths every year (www.cgs.gov.cn). With the rapid urbanization and economic development in China, there is an increasing risk of frequent and high magnitude landslide geohazard due to expansions of infrastructures and human settlements and serious degradations of the environment. Earthquake, heavy rainfall and human activities are the main triggers of landsliding. Especially, damages from earthquake-induced landslides may exceed damages from the earthquake itself (Fan et al. 2018). Therefore, the pre-identification of the landslide-susceptible areas proved a cost-effective way and an urgent task for reducing the socio-economic losses in seismic-prone regions. Although temporal and magnitude information of future landslide occurrences are not involved in the "susceptibility" term, landslide susceptibility map is very helpful for rational decisions in regional land management and infrastructure planning.

Landslide susceptibility mapping (LSM) remains a challenging task as landsliding is controlled by various factors, i.e., geology, geomorphology, topography and seismic activities, etc. To date, various methods have been developed and

Responsible Editor: Biswajeet Pradhan

✉ Suhua Zhou
suhua_zhou@hotmail.com

[1] College of Civil Engineering, Hunan University, Changsha 410082, Hunan, China

[2] Key Laboratory of Building Safety and Energy Efficiency of the Ministry of Education, Hunan University, Changsha 410082, China

[3] National Centre of Excellence in Geology, University of Peshawar, Peshawar, Pakistan

introduced for LSM, including knowledge-based methods, statistical methods and physically based methods. The pros and cons of these methods have been intensively reviewed in literature (Huang and Zhao 2018; Lee 2019; Reichenbach et al. 2018). Of these methods, the statistical approaches were most widely applied due to their suitability for LSM over broad areas and complex terrains (Lee 2019). Based on the expression of the statistical relationship between landslides and predicting factors, the statistical methods can be categorized as bivariate, multivariate and machine-learning-based (Reichenbach et al. 2018).

The bivariate analysis investigates the relationship between landslide predictive factors and landslide distribution one by one. Weights of each category within each factor were calculated based on landslide densities. Dozens of bivariate methods have been successfully applied in literatures (Shahabi et al. 2013; Sujatha and Rajamanickam 2015; Hong et al. 2018; Shrestha et al. 2017), such as frequency ratio (FR), weight of evidence (WoE) and information value (IV). Particularly, the Bayesian theory model, using the WoE method, has proved promising and very useful in LSM as its robustness and flexibility.

The multivariate analysis considers all independent predictive factors together to rate their contribution in causing landslides based on the presence and absence of historical landslides within a defined mapping unit. Logistic regression (LR) is a popular and widely adopted multivariate method in LSM at various scales (Wang et al. 2014). The LR has a number of attractive features: (1) it solves a complex problem by outputting a simple binary result representing the presence and absence of future landslides; (2) the independent input factors for LR can be nominal, numerical, categorical or any combination of them; and (3) the independent variable are not required to be normally distributed (Pham et al. 2016).

Literatures have shown that the effectiveness of statistical methods in LSM largely depends on the nature of the factor, sampling way of the data, and even the size of the data set (Schicker and Moon 2012; Kavzoglu et al. 2015). However, in practical LSM applications, such requirements are hard to be satisfied. Hence, machine-learning-based methods are increasingly preferred in LSM (Reichenbach et al. 2018) in most recent LSM studies, especially when the input data are not normally distributed (Arabameri et al. 2019a, b; Kavzoglu et al. 2015). Popular machine-learning-based LSM methods include support vector machine (SVM) (Fanos et al. 2020), radial basis function (RBF) (Javdanian and Pradhan 2019), convolutional neural network (CNN) (Jena et al. 2020) and so on. These machine-learning algorithms had shown great potential in mapping of geohazards like landslides, rockfalls, earth dams and gully erosion. As a representative machine-learning tool, SVM has shown a good capacity to solve highly complicated nonlinear problems (Tehrany et al. 2015; Huang and Zhao 2018; Pham et al. 2018). Overall, each method mentioned above had shown efficiency for LSM in many individual studies. The performance of an LSM method largely depends on nature, the scale and the data availability of the study area (Kavzoglu et al. 2015). Given that, several comparative studies have been proposed in recent years and can be found in (Schicker and Moon 2012; Kavzoglu et al. 2015; Pham et al. 2018). The results of comparative studies differ among cases. An accurate landslide susceptibility map would be of paramount significance for regional sustainable development by pre-identifying of landslide-prone regions. Nevertheless, selection of appropriate LSM methods is an exigent task since no method is always outperformed than others (Pham et al. 2016; Chapi et al. 2017; Khosravi et al. 2018). Hence, a detailed comparative evaluation of different LSMs for a given area will aid to assess their performances and to optimize the suitable methods.

After an extensive literature survey and contemplating the sensitivity of addressing problem to practical obstacles, three kinds of statistical LSM models such as WoE, which is bivariate, LR, which is multivariate, and SVM, which is a machine-learning-based algorithm were adopted in the present case study a seismic-prone region in China along with their performance's comparison. The Lushan County in Sichuan Province of China, a mountainous and inland county which had suffered heavily from seismic activities and induced landslides, was taken as study area for the GIS-based LSM using different statistical models. The objectives of the present study aimed to evaluate and compare the performance of these models and finally to produce an accurate landslide susceptibility map of the study area, which would benefit for regional sustainable development and landslide hazard mitigation. The outcome of this study with a reliable map would also provide a reference for the surroundings of the study area and other similar terrains.

Study area

For the present study, the southern part of Lushan County, Ya'an city, Sichuan Province in China, was selected as a study area. It is situated between 102.52° E and 103.11° E longitudes, and 30.01° N to 30.49° N latitudes (Fig. 1). Located at the western margin of the Sichuan basin and the southwest of the Longmenshan fault zones (LFZ), the study area is characterized by frequent seismic activities, and it was the worst-hit region of the Lushan Earthquake in 2013. The mountainous terrains account for up to 90% of the Lushan County, ranging from high in the northwest to low in the southeast. Frequent seismic activities and the mountainous terrains make Lushan County and its surroundings very susceptible to landsliding.

Geological units in the study area mainly range from the Upper Proterozoic Sinian to the Mesozoic Cretaceous. The main rocks are gabbro, pyroclastics, carbonates, sandstones, shale and dolomites. Due to abundant precipitation and plenty of sunlight, the majority of these rocks had weathered in

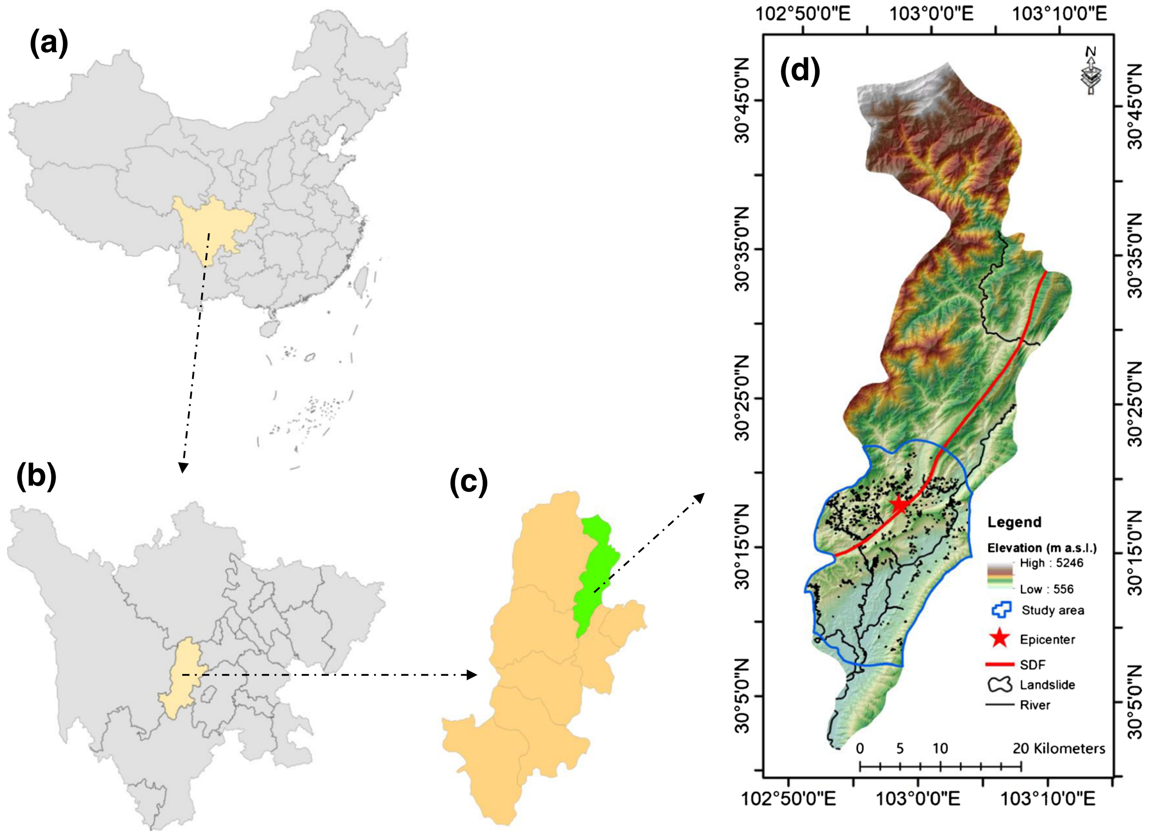

Fig. 1 Location of the study area: **a** China; **b** Sichuan; **c** Ya'an City and **d** Lushan County

varying degrees on the surface. In some regions, the weathering even penetrated deep into the rock masses through joints and structural planes. As a segment of the Longmenshan fault and the source fault of the 2013 Lushan event, the Shuangshi-Dachuan fault (SDF) is the main active fault in Lushan County. According to Chen et al. (2013), the Lushan event did not significantly reduce the seismic risk of the SDF. A high risk of potential earthquakes with a magnitude between Mw7.2 and Mw7.3 still exists in Lushan County and its surroundings. Therefore, it is of great importance to identify landslide-susceptible regions for the study area.

Materials

Landslide inventory

A landslide inventory registers the location, occurring time and characteristics of past landslides. It not only plays an indispensable role in LSM but also benefits seismic hazard mitigations. Until now, compiling of a detailed landslide inventory remains a challenging task. Generally, there were two kinds of landslide inventory. One is the historical inventory, recording all the landslides that happened within an historical period and regardless of their triggers. The other one is the event inventory, which portrays all the landslides triggered by a specific trigger (e.g., an earthquake or an intensive rainfall). Since no historical inventory was available and seismic is a dominant inducing factor for landslides in the study area, an event inventory of landslides induced during the Lushan earthquake was used in this study.

As an outcome of emergency response to the devastating Lushan earthquake, the adopted inventory containing 942 landslides was produced using high-resolution aerial photos and satellite images by the Chinese Academy of Science (CAS) (Fig. 2). The boundaries of these landslides were delineated and mapped with the help of GIS. According to the post-seismic field investigations of some landslides, they are of three types, namely, translational landslides, rotational landslides, rockfalls and debris flows. Landslide mechanism differs in each type. Even so, it is impossible to check the type of all landslides, as their locations are hard to be reached. Therefore, all landslides were considered in this present study.

Landslide predictive factors

Landslide occurrence is a comprehensive consequence of geological and environmental factors. Hence, the first step in performing LSM using statistical methods is to select landslide predictive factors. There are no universal rules for the selection of landslide predictive factors. Different researchers have applied different sets of factors in LSMs (Reichenbach et al. 2018; Lee 2019).

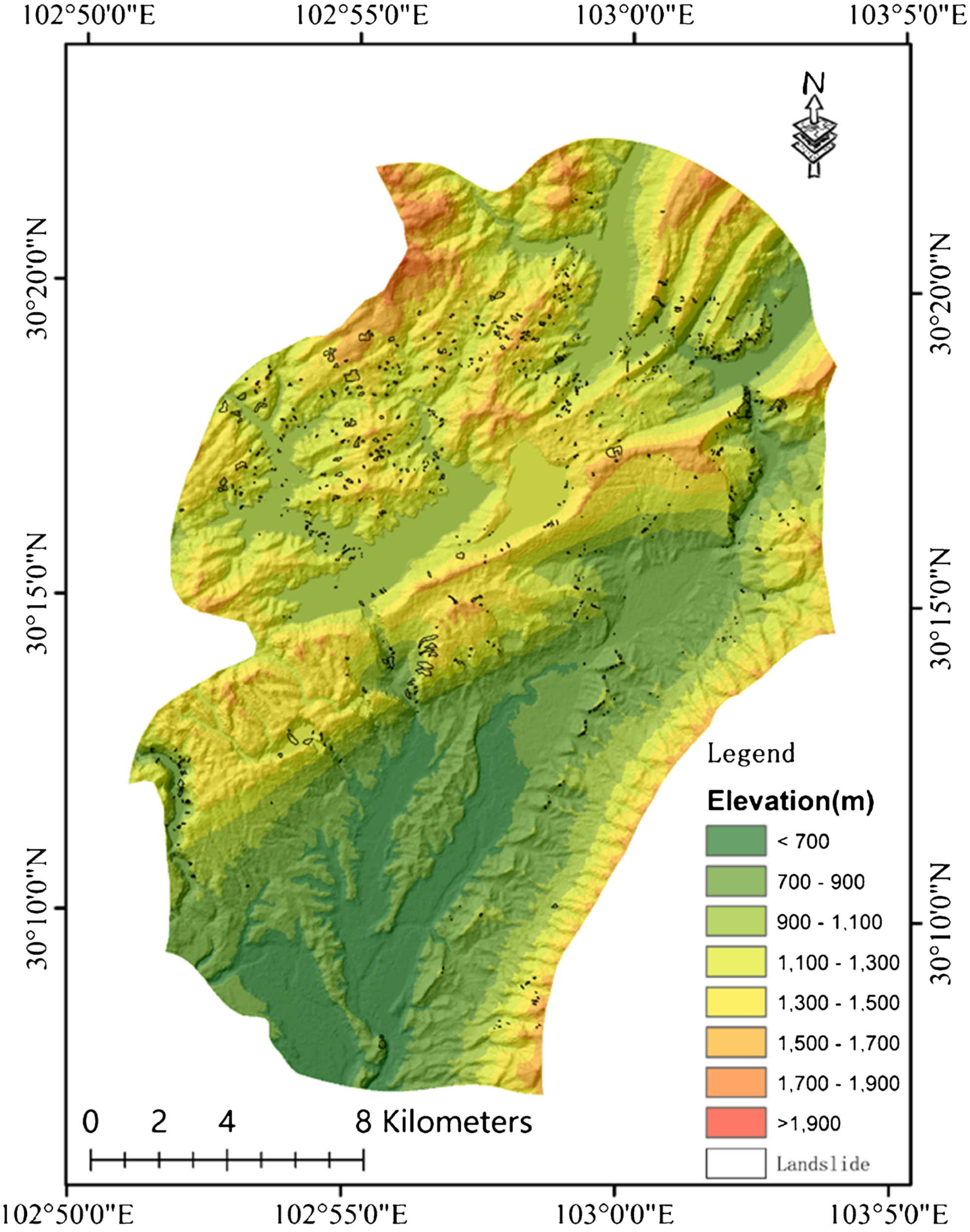

Fig. 2 Map showing the distribution of landslides in the study area

Despite this situation, it has been widely accepted that data availability, analysis scale and characteristics of the study area determine the set of landslide predictive factors. In this study, landslide predictive factors were selected based on the following criteria: (1) data of predictive factors available from reliable organizations or government; (2) resolution of data is sufficient to facilitate the LSMs; (3) digitalized format is preferred in order to minimize the errors. Based on the above considerations, ten landslide predictive factors including four topographic factors (including elevation, slope, slope aspect and curvature), three hydrological factors (including distance to the river, topographic wetness index and stream power index), the geological factor, the landcover factor and the seismic factor (PGA) have been used in this study (Table 1). A brief introduction to these factors is given in the following sections.

Topographic factors

Several topographic factors can be derived from a digital elevation model (DEM). In this study, the DEM with a spatial resolution of 12.5 m was a product of ALOS PALSAR, which is freely available at Alaska Satellite Facility Distributed Active Archive Centre (ASF DAAC) (www.asf.alaska.edu). As illustrated in Fig. 3a, the study area ranges from 610 to 1951 m in elevation. Using a 200-m interval, the elevation factor was divided into 8 subsets: (1) <700 m; (2) 700–900

Table 1 List of landslide predictive factors

Factors		Format	Source
Topographic factors	Elevation	Raster	Derivation of DEM (DEM is freely available at Alaska Satellite Facility Distributed Active Archive Centre)
	Slope	Raster	
	Aspect	Raster	
	Curvature	Raster	
Hydrological factors	DtoR	Raster	National Geology Library of China
	TWI	Raster	Secondly derivation of DEM
	SPI	Raster	
Geological factor	Lithology	Shapefile	National Geology Library of China
Land cover	PTC	Raster	Geospatial Information Authority of Japan
Seismic factor	PGA	Shapefile	USGS

m; (3) 900–1100 m; (4) 1100–1300 m; (5) 1300–1500 m; (6) 1500–1700 m; (7) 1700–1900 m; and (8) >1900 m.

Using a DEM, factors of slope gradient, aspect and curvature were produced. The slope gradient is a direct driver for landslides. The possibilities of landsliding become high as the slope gradient increases. The slope gradient of the study area was calculated from a first derivate of the DEM, which ranges between 0 and 82°. Using a 10° interval, slope gradient was divided into seven categories: (1) 0–10°; (2) 10–20°; (3) 20–30°; (4) 30–40°; (5) 40–50°; (6) 50–60°; (7) 60–70°; (8) 70–82° (Fig. 3b). The slope aspect describes the facing direction of the slope surface. Vegetation cover and degree of weathering of the slope materials vary in different slope aspects, due to the differences in precipitation, temperature and wind. In this study, the slope aspects were categorized as follows: (1) north (N); (2) northeast (NE); (3) east (E); (4) southeast (SE); (5) south (S); (6) southwest (SW); (7) west (W); (8) northwest (NW); and (9) flat (F) (Fig. 3c). Curvature is a second-order derivative of the DEM. A positive curvature value indicates an outwards convex surface, while a negative value means an inwards concave surface. A curvature value near zero (e.g., between −1.0 and 1.0) means that the slope is close to flat. Three general categories of curvature factor were

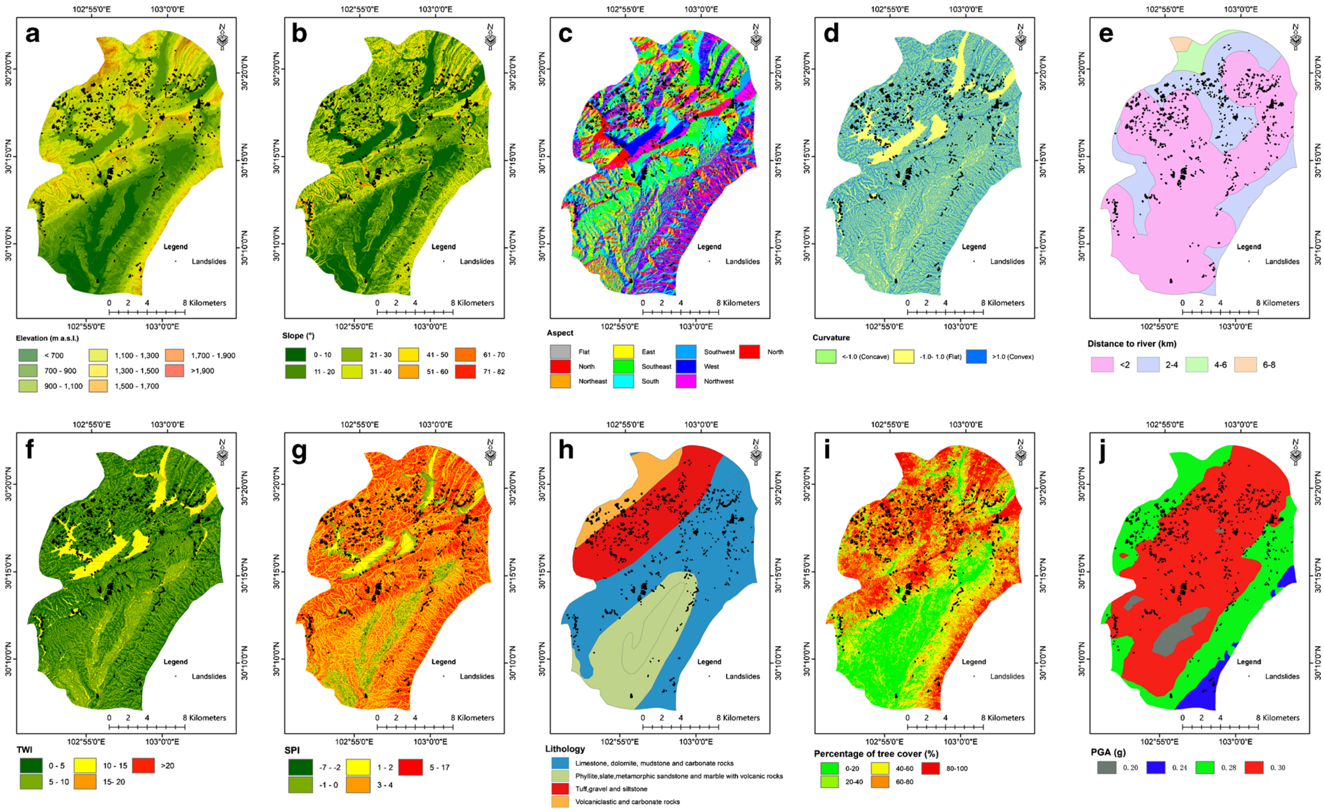

Fig. 3 Landslide predictive factors: **a** elevation; **b** slope gradient; **c** aspect; **d** curvature; **e** distance to the river; **f** TWI; **g** SPI; **h** lithology; **i** percentage of tree cover; and **j** PGA

defined herein: (1) < −1.0 (concave); (2) −1.0–1.0 (flat); (3) >1.0 (convex) (Fig. 3d).

Hydrologic factors

Hydrological factors play an important role in the regional landscape evolution. Landslides are usually distributed along drainages because erosion and excavation near rivers provide the accumulation of the necessary potential energy and the free surface for landslide occurrences. Hence, the factor of distance to the river (DtoR) was taken into account to assess the correlations between landslides and rivers. The DtoR was reclassified into four categories using a 2-km distance interval using the buffer function in GIS: (1) 0–2 km; (2) 2–4 km; (3) 4–6 km; and (4) 6–8 km (Fig. 3e).

The topographic wetness index (TWI) and the stream power index (SPI) are two widely used factors modelling the runoff process. The TWI indicates the conditions of soil moisture, groundwater flowing and accumulation. The TWI of the study area ranges from 0 to 23.83. As shown in Fig. 3f, the TWI factor was classified into five categories: (1) 0.0–5.0; (2) 5.0–10.0; (3) 10.0–15.0; (4) 15.0–20.0; and (5) >20.0. The SPI is another topographic factor measuring the erosive power of flowing water. The SPI values within the study area were arranged in four classes: (1) −7 to −2; (2) −1–0; (3) 1–2; (4) 3–4; (5) 5–17 (Fig. 3g). Both TWI and SPI can be calculated using slope gradient and catchment area. Built on the steady-state assumption, for the uniform soil condition, the TWI and the SPI can be calculated using Eq. 1 and Eq. 2, respectively, as follows (Sorensen et al. 2006):

$$\mathrm{TWI} = ln(A_s/\tan\beta) \quad (1)$$

$$\mathrm{SPI} = \ln(A_S \cdot \tan\beta) \quad (2)$$

where A_S is the upslope contributing area per unit length of contour (m^2/m) of the cell, and β is the slope gradient of the cell.

Geological factors

Lithology of rocks is among the deterministic factors of landslides in literature. Differences in lithology lead to the variance of strength, permeability and structure development. The 1:200,000 geology map in digital GIS shapefile format was collected from the National Geology Library of China. Within the study area, four lithological groups were included, namely, the Guankou group, the Baitianba group, the Guanwushan group and the Lushan group (Fig. 3h). The Guankou group mainly consists of limestone, dolomite, mudstone and carbonate rocks. The Baitianba group is composed of tuff, gravel and siltstone. The Guanwushan group constituted volcaniclastic and carbonate rocks. In the case of the Lushan group, phyllite, slate, metamorphic sandstone and marble with volcanic rocks are the dominant lithology rocks.

Percentage of tree cover factor

Bare ground is more prone to landsliding than ground with vegetation or tree cover. Therefore, the tree cover percentage was taken into account as a predictive factor related to the landslide occurrence. Percentage of tree cover (PTC) is a quantitative parameter, defined as the ratio of surface area with branches and leaves of trees coverage to the total surface area. In this study, the PTC factor was derived from the Geospatial Information Authority of Japan. Five classes of PTC were set in this study: (1) 0–20%; (2) 20–40%; (3) 40–60%; (4) 60–80%; and (5) 80–100% (Fig. 3i).

Seismic factor

Landslides were triggered by strong ground shaking on slope masses. Peak ground acceleration (PGA) represents the leading indicator of seismic scenarios, measuring the maximum shaking intensity at a given ground point. The parameter of PGA is widely used by engineers for the purpose of seismic resistance designing. In the field of seismic hazard mitigation, PGA maps of a particular seismic event have been very helpful in damage assessment at a regional scale. The regional PGA data of the Lushan earthquake was obtained from the USGS ShakeMap (http://earthquake.usgs.gov) in GIS shapefile format. The highest PGA value of 0.3 g occurred near the epicenter of the mainshock. In the USGS ShakeMap, the study area was reclassified into four categories according to the PGA values as (1) 0.2 g; (2) 0.24 g; (3) 0.28 g, and (4) 0.30 g (Fig. 3j).

Methods

Flowchart of LSM

Even for the same dataset, the results of LSM using different models may vary. Literature reviews have shown that it is hard to tell which model outperformed the others (Huang and Zhao 2018; Reichenbach et al. 2018; Lee 2019). Hence, the efficiency and accuracy of different LSM models should be evaluated and compared in order to obtain a reliable landslide susceptibility map. This paper demonstrates the application and comparison of the WoE, the LR and the SVM in generating landslide susceptibility maps of a selected area affected by the Lushan earthquake. The theoretical background and implementation of these models is briefly introduced in this section.

The general flowchart of conducting the LSMs is shown in Fig. 4. Firstly, the landslide inventory and all the landslide predictive factors were rasterized into raster of 12.5 m. The landslide raster was randomly divided into two parts. 70% of

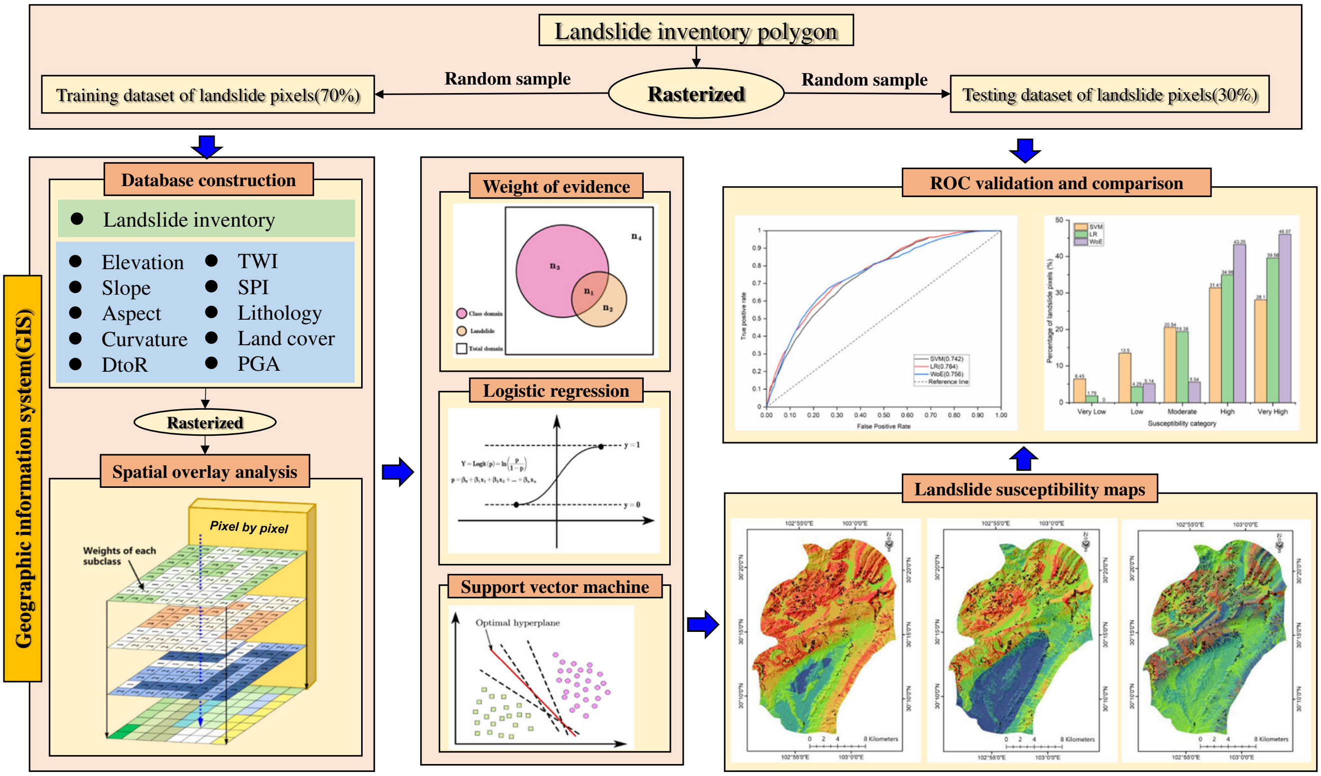

Fig. 4 Flowchart of LSM in this study

landslide raster cells were used to develop or train the model. 30% of landslide raster cells were used to validate the results. Then, each of the LSM models (WoE, LR and the SVM) was applied individually to produce the landslide susceptibility map. After that, accuracies of resultant maps were validated and compared.

Generation of training and testing datasets

For present study, the identified 867 landslides polygons were first rasterized into 18,712 pixels having a spatial resolution of 12.5 m using ArcGIS10.2. Then, the whole dataset was randomly split into 2 subsets: (i) the training dataset containing 13,099 pixels (70% of the total) and (ii) the validation dataset containing 5613 pixels (30% of the total). For the WoE model, only the landslide dataset is required as input. However, both the landslide and non-landslide datasets are required in the LR and SVM modelling. In order to generate the non-landslide dataset, an assumption was made in this study; the pixels located 50 m (4 times of the raster resolution) away from the known landslide pixel were considered as non-landslide regions, because the pixel approximating the landslides display similar conditions as the landslide pixel and might cause problems in building the prediction model. Finally, the same counts of non-landslide pixels (18,712 pixels) were randomly selected in the non-landslide regions. Hence, for the WoE model, there were only 13,099 landslide pixels in the training dataset. As a comparison, there were 13,099 landslide pixels and 13,099 non-landslide pixels in the training dataset of the LR and the SVM. For the validation dataset of all the three models, there were 5613 landslide pixels and 5613 non-landslide pixels.

LSM models

Weight of evidence

Based on the Bayesian theory of conditional probability, the weight of evidence (WoE) is among the most popular bivariate statistical models for quantification of spatial correlations between evidence (landslide-predicting factors) and known landslide occurrences. More detailed information about the theory and formulations of WoE can be found in Bonham-Carter (1994). Derived from the Bayesian rules, the weight of each landslide-predicting factor is calculated as the posterior probability of landsliding given the presence or absence of the landslide-predicting factors, the WoE can be mathematically defined as follows:

$$\begin{cases} w_i^+ = \ln \dfrac{P\{B_i|L\}}{P\left\{B_i\middle|\overline{L}\right\}} \\ w_i^- = \ln \dfrac{P\left\{\overline{B}_i\middle|L\right\}}{P\left\{\overline{B}_i\middle|\overline{L}\right\}} \end{cases} \tag{3}$$

where P is the conditional probability; B_i and $\overline{B}_i$ denote the existence and absence of a landslide-predicting factor, respectively. L and $\overline{L}$ are the existence and absence of a landslide. For each factor, the positive weight w^+ is used to measure the importance of factor existence for landsliding. A $w^+ > 0$ indicates the factor's existence is favourable for landsliding, while a $w^+ < 0$ means it is unfavourable for landsliding. The negative weight w^- evaluates the importance of the factor's absence for landsliding. When $w^- > 0$, the absence of the factor is favourable for landsliding, and when $w^- < 0$, it is unfavourable for landsliding.

The overall correlations between a landslide predictive factor and the landslides occurrences can be measured using the magnitude of contrast weight (C) defined as the difference between the positive weight and the negative weight. The contrast can be calculated as follows:

$$C = w_i^+ - w_i^- \tag{4}$$

The primary WoE was performed for the binary factors, which include only two classes (absence and presence of the factor). However, most of the factors in the LSM were multiclass factors (e.g., slope gradient). By writing Eq. 3 in the number of GIS raster pixels, the WoE was performed with each landslide predictive factor in a more practical way to calculate w^+ and w^-. For each landslide predictive factor, the weights of each class can be obtained using Eq. 5:

$$\begin{cases} w^+ = \ln\left(\dfrac{n_1/(n_1+n_2)}{n_3/(n_3+n_4)}\right) \\ w^- = \ln\left(\dfrac{n_2/(n_1+n_2)}{n_4/(n_3+n_4)}\right) \end{cases} \tag{5}$$

where n_1 is the count of landslide pixels in the current class, n_2 is the count of landslide pixels not in the current class, n_3 is the count of non-landslides pixels in the current class and n_4 is the number of non-landslide pixels not in the current class.

Finally, all of the conditional independent landslide predictive factors were spatially integrated to calculate the landslide susceptibility index (LSI) of each pixel as follows:

$$\mathrm{LSI}_{(x,y)} = \sum_{i=1}^{n} C_{x,y}^{i} \tag{6}$$

where (x, y) denotes the spatial location of each pixel and n is the total number of factors.

Logistic regression

The logistic regression (LR) is a representative multivariate statistical method. It assesses the correlations between one dependent variable (landsliding) and several independent variables (predictive factors). One of the distinguishing features of the LR is that the types of landslide predictive factors can be nominal, discrete, continuous or any mix of them. In a common LR modelling, the dependent variable has only two values ("1" and "0"). Through a link function, the dependent variable of the usual linear regression model was transformed into a logit variable in the LR. The probability of the dependent variable showing the value of "1" was estimated using the maximum likelihood estimation.

For the LSM using LR, it aims to optimize a best-fitting model to interpret the correlations between the landslides and their predictive factors. The dependent variable was coded by two values, "1" indicating the occurrence of a landslide, while "0" indicating its absence. The landslide predictive factors can be measured as follows: the nominal (e.g., slope aspect), the continuous (e.g., slope gradient, altitude) and the discrete (e.g., seismic intensity). The probability of landslide occurrence can be calculated through LR modelling, using Eq. 7 and Eq. 8:

$$Y = \mathrm{Logit}(p) = ln\left(\frac{p}{1-p}\right) \tag{7}$$

$$p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \tag{8}$$

where p is the probability when the value of the dependent variable Y is 1, n is the count of factors, β_0 is the intercept parameters in linear regression and β_i (i = 1, 2, …, n) is the regression coefficient, which can be used to measure the influence of each factor x_i (i = 1, 2, …, n) on landsliding.

Support vector machine

The support vector machine (SVM) has shown great potential in solving the classification and regressions problems. As a supervised machine-learning algorithm, a detailed description of SVM algorithms can be found in Cortes and Vapnik (1995) and more recently in Kumar et al. (2017). The general algorithm of the SVM can be summarized as follows: Given a linearly separable problem, a set of vectors $V = \{(\overrightarrow{x}_1, y_1), (\overrightarrow{x}_2, y_2), \ldots, (\overrightarrow{x}_n, y_n)\}$, has n points $\overrightarrow{x}_i$, with two labels (y = ±1) indicating the two classes the $\overrightarrow{x}_i$ belonging to. The principle of the SVM is to find the optimized decision boundary separating $\overrightarrow{x}_i$ belonging to the group $y = +1$ from the $\overrightarrow{x}_i$ for $y = -1$ with a maximum margin. An optimized decision boundary can be mathematically written as:

$$\overrightarrow{w} \cdot \overrightarrow{x} - b = 0 \tag{9}$$

where $\overrightarrow{w}$ is the normal vector of the hyperplane. $\frac{b}{\|\overrightarrow{w}\|}$ is the offset of the hyperplane from the origin along the direction of $\overrightarrow{w}$. Geometrically, the distance between the two hyperplanes is $\frac{2}{\|\overrightarrow{w}\|}$. To separate the two classes of points $\overrightarrow{x}_i$, there are many possible hyperplanes that can be chosen. Optimal classification occurs when a hyperplane provides a maximum distance to the nearest training data points. Thus, the classification can be stated as a constrained optimization problem:

$$\underset{w,b}{\text{Minimize}} \quad \left\| \overrightarrow{w} \right\| \tag{10}$$

Subject to:

$$y_i\left(\overrightarrow{w}\cdot\overrightarrow{x}-b\right)\geq 1-\xi_i, \qquad i=1,2,\ldots,n \tag{11}$$

For the non-separable problems, a slack variable ξ_i was introduced in the original constrains (Eq. 8) to allow for the error tolerance as follows:

$$y_i\left(\overrightarrow{w}\cdot\overrightarrow{x}-b\right)\geq 1-\xi_i, \qquad i=1,2,\ldots,n \tag{12}$$

By introducing a Lagrange multiplier, the cost function can be mathematically expressed as:

$$\text{Minimize} \sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\left(x_i x_j\right) \tag{13}$$

Subject to:

$$\sum_{i=1}^{n}\alpha_i y_i=0, \qquad 0\leq\alpha_i\leq C \tag{14}$$

where $\alpha=(\alpha_1, \alpha_2, \ldots, \alpha_n)\in R^n$ is the Lagrange multiplier and C is the penalty term.

The cost function can be written as:

$$g(x)=\text{sign}\left(\sum_{i=1}^{n}y_i\alpha_i x_i+b\right) \tag{15}$$

In some cases, when there existed no linear hyperplane separating the input vectors, some kernel functions were used to project the input data into a higher dimensional feature space. The cost function is rewritten as:

$$g(x)=\text{sign}\left(\sum_{i=1}^{n}y_i\alpha_i k\left(x_i,x_j\right)+b\right) \tag{16}$$

where $k(x_i, x_j)$ is the kernel function. The mathematical formulations of the four most frequently used kernels in SVM are listed in Table 2.

Validation of the accuracy

For LSM, a landslide pixel was labelled as "positive", while a non-landslide pixel was labelled as "negative". If the predicted label of any pixel agrees with its true label, a prefix "true" was added to the label of the pixel. On the contrary, if the predicted label of any pixel disagrees with its true label, a prefix "false" was added to the label of the pixel. Given this, any pixel can be labelled as "true positive" (TP), "true negative" (TN), "false positive" (FP) or "false negative" (FN) according to its true state and predicted condition (Table 3). The false positive rate (FPR) refers to the proportion of actual negative but wrongly predicted as positive cases among the total positive cases (Eq. 17). The true positive rate (TPR) refers to the proportion of actual positive and correctly predicted as positive cases among the total positive cases (Eq. 18).

$$\text{FPR}=\frac{\text{FP}}{\text{FP}+\text{TN}} \tag{17}$$

$$\text{TPR}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{18}$$

Accuracy validation was the final and indispensable step in LSM. The receiver operating characteristic (ROC) analysis is the most popular validation method for classification. The ROC analysis was performed by drawing a curve on an *x*-*y*-plane, with the horizontal axis showing FPR and the vertical axis showing TPR. For LSM, the predicted output for each landslide pixel is an LSI value, which is continuous rather than a binary result. In order to obtain the ROC curve, a series of threshold values "*T*" was set to define the predicted labels of each pixel. If LSI> *T*, then it was predicted as "positive" otherwise "negative". For each "*T*", the corresponding pair of (TPR, FPR) can be calculated and plotted as a point on the ROC plane. The ROC curve was then obtained by connecting all the points. For the present study, the training dataset and the validation dataset were analysed using ROC curve, respectively.

The area under the ROC curve (AUC) was a quantitative index for the overall performance evaluation of the LSM models. Typically, the AUC value varies between 0.5 and 1. If the AUC value of a classification model is 0.5, the classification model performed no better than chance. A higher AUC value indicates a better performance of a classification model. An AUC value of 1.0 is provided by a perfect classification model.

Implementation of LSM models

Implementation of the WoE

As shown in Table 4, the weight of each landslide predictive factor category was calculated using Eq. 6 by overlaying the 13,099 landslide pixels and the landslide predictive factors. In order to measure the contribution magnitude of each predictive factor, the contrast values were also calculated using Eq. 4.

The conditional independence (CI) between the landslide predictive factors used in WoE should be evaluated before they can be used to produce the LS map using WoE. In this study, the CI was tested using the Chi-square method. Based on the positive or negative contrast values of each category,

Table 2 Four commonly used kernel functions

	Kernel function	Formulation*	Kernel parameters
1	Linear (LN)	$k(x_i\ , x_j) = x_i^T x_j$	
2	Polynomial (PL)	$k(x_i\ , x_j) = \left(\gamma x_i^T x_i + r\right)^{\ d}$	γ, d
3	Radial basis functions (RBF)	$k(x_i,x_j)=\exp(-\gamma\|x_i-x_j\|^2)$	γ
4	Sigmoid (SIG)	$k(x_i\ , x_j) = \tanh\left(\gamma x_i^T x_j + r\right)$	γ

* γ is gamma term in all kernels except the LN; d is the polynomial degree; r is the bias term

each landslide predictive factor was first converted into a binary pattern. Then, x^2 value for each pair of 10 binary factors was calculated at the 99% confidence level and 1 degree of freedom. An x^2 value less than 6.635 indicates that the pairs were conditionally independent in predicting landslides. As shown in Table 5, 18 of the total 45 pairs of landslide predictive factors were conditionally independent (CI). Therefore, two potential combinations of factors can be used in the LSM (Table 6). Since slope gradient is an indispensable controlling factor of landslide occurrences, combination II was finally used in this study.

Implementation of the LR

A database containing 26,198 cells was created. Each cell in the database has 10 attributes representing the category value of each factor and a binary label value of 1 and 0 (existence and absence of landslide). In LR modelling, the attributes were used as independent variables, while the binary labels were used as the dependent ones. For the current study, the binary LR statistical analysis was performed using Python. The correlation between landslides and predictive factors is shown in Table 7, where the multicollinearity of independent variables was examined using the tolerance (TOL) and variance inflation variables (VIF).

Since not all the independent factors contribute significantly to landsliding, a forward stepwise LR was carried out to select the factors closely related to landslide occurrences. In the beginning of this type of regression, only one variable was included in this model. Then for each step followed, independent variables were added in the regression one by one after evaluation. During each step, the regression model with an independent variable added was evaluated using the maximum likelihood ratio (MLR). If the change in the logarithm of likelihood was less than the probability of factors kept in the model, the variable was entered in the regression. Otherwise, it was excluded. After repeating this process for each factor, a regression model including all significant independent variables was built, and the coefficients were assigned to each independent variable. In the present study, the threshold probability for variables to be included in the regression model was set to 0.05.

Implementation of the SVM

The SVM modelling was performed through Python programming in this study. The free python machine-learning python package "scikit-learn" (Pedregosa et al. 2011) was used to facilitate the modelling. Four common types of kernel functions mentioned in Table 2 are given in the "scikit-learn". The LN is a case of the PL, when $\gamma = 1$ and $d = 1$. Despite its good extrapolation abilities in lower-order fitting, "overfitting" failure may occur if the order of a polynomial is high while using the PL. The SIG are derived from neural networks and are widely used in deep learning. The RBF has good interpolation abilities in mapping a sample into a higher dimensional space. The RBF performs well in both large and small samples and has fewer parameters as compared to the polynomial kernel function. Previous studies of LSM using SVM have indicated that in most cases, the RBF works best among the four types of kernels (Tehrany et al. 2015; Huang and Zhao 2018). Therefore, the RBF kernel was used in this study.

The accuracy of the SVM using kernel was largely dependent on the kernel parameters. In the SVM modelling with

Table 3 Confusion matrix for the ROC analysis

		Predicted label	
		Landslide (positive)	Non-landslide (negative)
Actual label	Landslide (positive)	TP[1]	FN[2]
	Non-landslide (negative)	FP[3]	TN[4]

[1] *TP*, no. of the actual positive and correctly predicted cases

[2] *FN*, no. of the actual positive but wrongly predicted cases

[3] *FP*, no. of the actual negative but wrongly predicted cases

[4] *TN*, no. of the actual negative but correctly predicted cases

Table 4 Results of the WoE analysis

Factor	Class	No. of landslide pixels	No. of total pixels	w^+	w^-	C
Aspect	North	1357	282,347	−0.088	0.011	−0.099
	Northeast	911	199,741	−0.141	0.011	−0.152
	East	2071	259,355	0.423	−0.063	0.485
	Southeast	1604	421,118	−0.322	0.054	−0.376
	South	2307	341,578	0.254	−0.047	0.301
	Southwest	1296	297,453	−0.187	0.023	−0.210
	West	1832	353,333	−0.012	0.002	−0.014
	Northwest	1720	341,160	−0.040	0.006	−0.046
Elevation (m)	<700	50	249,104	−3.269	0.102	−3.371
	700–900	1271	541,233	−0.807	0.143	−0.950
	900–1100	2657	462,820	0.090	−0.022	0.112
	1100–1300	4662	569,189	0.448	−0.182	0.630
	1300–1500	3807	523,774	0.328	−0.108	0.436
	1500–1700	652	131,533	−0.057	0.003	−0.060
	1700-1900	0	18,185	0.000	0.007	0.000
	>1900	0	247	0.000	0.000	0.000
Slope (°)	<10	1400	605,554	−0.823	0.166	−0.989
	10–20	1614	602,170	−0.675	0.145	−0.820
	20–30	3318	682,483	−0.077	0.027	−0.104
	30–40	2807	375,862	0.355	−0.078	0.433
	40–50	2164	176,524	0.855	−0.108	0.963
	50–60	1093	39,582	1.683	−0.071	1.755
	60–70	507	12,838	2.053	−0.034	2.088
	70–82	197	1072	3.754	−0.015	3.768
Curvature	<−1.0	7453	716,572	0.689	−0.505	1.195
	−1.0–1.0	476	121,621	−0.295	0.013	−0.308
	>1.0	5170	1,658,168	−0.523	0.594	−1.116
DtoR (km)	<2	9435	1,763,753	0.019	−0.048	0.068
	2–4	3616	651,300	0.057	−0.021	0.078
	4–6	47	63,743	−1.967	0.022	−1.989
	6–8	0	17,565	0.000	0.007	0.000
Lithology	Group A	6301	1,174,740	0.022	−0.020	0.042
	Group B	4447	548,203	0.439	−0.168	0.606
	Group C	1768	172,355	0.675	−0.074	0.749
	Group D	582	601,063	−1.694	0.231	−1.926
PGA	0.2	7	92,329	−4.242	0.037	−4.280
	0.24	106	98,422	−1.588	0.032	−1.620
	0.28	4032	780,915	−0.016	0.007	−0.023
	0.30	8953	1,524,695	0.113	−0.208	0.321
SPI	−7.0 to −2.0	24	27,315	−1.792	0.009	−1.801
	−2.0–0.0	440	158,451	−0.639	0.032	−0.671
	0.0–2.0	1720	529,837	−0.482	0.098	−0.581
	2.0–4.0	5139	1,054,266	−0.074	0.051	−0.125
	>4.0	5775	725,749	0.419	−0.239	0.658
PTC (%)	<20	1086	538,873	−0.960	0.158	−1.118
	20–40	717	222,386	−0.489	0.037	−0.526
	40–60	3695	562,560	0.226	−0.076	0.302
	60–80	4861	674,743	0.319	−0.149	0.468
	>80	2740	497,796	0.048	−0.012	0.060

Table 4 (continued)

Factor	Class	No. of landslide pixels	No. of total pixels	w^+	w^-	C
TWI	0.0–5.0	4318	1,073,334	−0.267	0.163	−0.430
	5.0–10.0	6198	1,143,291	0.032	−0.028	0.061
	10.0–15.0	2461	256,292	0.608	−0.100	0.709
	15.0–20.0	81	18,659	−0.191	0.001	−0.192
	>20.0	41	4042	0.664	−0.002	0.665

RBF kernel, the γ and the penalty term C need to be optimized. The γ determines the nonlinearity degree in SVM, while the penalty term C was used to avoid overfitting of SVM by controlling the trade-off between training errors and the margin. A grid search with $C = 2^{-10}, 2^{-9}, \ldots, 2^{10}$ and $\gamma = 2^{-10}, 2^{-9}, \ldots, 2^{10}$ was carried out to search the best pairs of (C, γ). For each pair, a multi-folder cross-validation method (MFCV) was adopted in this study (Chen et al. 2017).

Results

Correlations between predictive factors and landslide distribution

Landslides distributed unevenly within the study area. As shown in Fig. 5a, the majority of landslides occurred in area between 700 and 1500 m in elevation. The category of 1100–1300 m was most vulnerable for landsliding since its 22.8% area coverage accounts for 35.6% landslide occurrences. For the slope gradient, the highest percentage of landslides occupation (25.3%) was found in the category of 20–30°, while the category of 60–82° accounts for the lowest landslide percentage of 27.3% (Fig. 5b). In the case of slope aspect, no general correlation has been found between landslides and aspects. Landslides mostly concentrated on south and east-facing slopes (Fig. 5c). The result has indicated that slopes with the inwards concave surface (56.9%) were more favourable for landslides than that with an outwards convex surface (39.5%), as 28.7% of landslides occupied in the former and 66.4% of them occupied in the latter (Fig. 5d).

As shown in Fig. 5e, 72.3% of the landslides concentrated within 2 km buffer of rivers. Moreover, the concentration decreases as the distance increases. In general, the high saturation level favours the landslide initiation by weakening the shear strength of slope masses and increasing the sliding force. As can be seen from Fig. 5f, 47.32% of the landslides occurred in the area with the TWI value of 5.0–10.0, and 18.8% of total landslides occurred in the category of 10.0–15.0. In the case of the SPI, possibilities of landslide occurrences are greater in the high value of stream power. 39.23% and 44.09% of the landslides occurred in the SPI category of 2.0–4.0 and >4.0, respectively (Fig. 5g). Landslide susceptibility levels varied in different lithology of rocks.

For the case of lithological rocks, the Guankou group accounts for 48.1% of total landslides. Followed by the Baitianba group, 34% of the landslides occupied this group. As can be seen from Fig. 3a, b, coverages of these two lithological groups were dominated by mountainous terrains with a higher elevation and steeper slopes. In addition, due to plentiful precipitation, sufficient illumination and frequent seismic events within the study area, most of the lithological rocks have undergone some certain degree of weathering, making slopes more vulnerable to landsliding. As a comparison, only

Table 5 Conditional independence test using pairwise chi-square statistics

Factors	Elevation	Slope	Curvature	DtoR*	Lithology	PGA	SPI	PTC	TWI
Aspect	**2.56**	106.93	65.39	112.36	**0.85**	99.51	550.21	105.28	462.53
Elevation		**4.53**	322.42	**5.76**	47.14	**1.65**	**5.08**	**2.54**	139.65
Slope			198.72	**1.23**	39.76	**2.32**	**4.65**	15.41	76.51
Curvature				115.82	**1.23**	181.92	95.24	278.32	**5.79**
DtoR*					36.79	**4.81**	**4.67**	**0.61**	112.51
Lithology						550.62	220.18	421.74	46.86
PGA							**5.3**	**0.51**	337.69
SPI								**2.12**	93.03
PTC									64.47

* DtoR: distance to the river

The bold values indicate that the pairs are conditionally independent

Table 6 Possible combinations of conditionally independent factors used in WoE

Combination type	Factors included
Combination I	Elevation, slope, distance to the river, PGA, SPI
Combination II	Elevation, distance to the river, PGA, SPI, PTC

4.44% of landslides occurred in the Lushan group, which accounts for 24% of the total area (Fig. 5h).

Distribution of landslides has shown that the area with tree cover of 60%–80% accounted for most landslide occurrences (37.1%), while the PTC of 20–40% had the lowest percentage of landslides (5.47%) (Fig. 5i). For the factor of PGA, the areas suffering from stronger ground shaking are more vulnerable to landsliding. 68.35% of the landslides concentrated within the PGA category of 0.30 g (Fig. 5j).

Results of the WoE

The landslide susceptibility map produced using the WoE is illustrated in Fig. 6a and the results of the WoE analysis are summarized in Table 2. The magnitude of spatial associations between landslide occurrences and each category of each landslide predictive factor was measured using the contrast values. A positive contrast value indicates a positive spatial association. The higher the contrast value is, the more positive the association is.

As illustrated in Table 2, east-facing slopes are most susceptible for landsliding, while the southeast-facing slopes have the lowest contrast value. In the case of elevation, the landslides were the most abundant in the category of 1100–1300 m. As for slope, steeper terrains were more prone to landsliding. The contrast values increase as the slope gradient increases. The category >60 had the highest contrast value of 1.986. In terms of curvature, the contrast values for concave, flat and convex regions were 0.501, −0.678 and −1.659, respectively. For the lithology, the highest contrast value was observed in the category of volcaniclastic and carbonate rocks (the Guanwushan Group). In the case of distance to the river, no category had a positive contrast value. The lowest contrast value of −2.351 was observed in the category of 2–4 km away from the river. Contrast values of all PGA categories were also negative. The category of 0.30 g had the lowest contrast value of −0.541. However, as the PGA increased, the contrast value got higher. For the factor of PCT, the category of <20% showed the lowest contrast of −1.504, while the category 60–80% had the highest value of −0.054. In terms of SPI, the more positive the SPI value, the higher is the contrast. As for the TWI, the highest contrast value of 0.709 was observed in the category of 10.0–15.0.

Table 7 The multicollinearity check among landslide predictive factors

Factors	TOL	VIF
Aspect	0.987	1.013
Elevation	0.452	2.212
Slope	0.319	3.136
Curvature	0.837	1.194
DtoR	0.722	1.384
Lithology	0.774	1.291
PGA	0.945	1.058
SPI	0.502	1.994
PTC	0.556	1.798
TWI	0.438	2.285

As previously stated, only conditionally independent factors were finally included in the WoE model to create the landslide susceptibility map. According to the CI test result (Table 5), the landslide susceptibility map was produced using elevation, slope, distance to the river, PGA and SPI. The weight of each category within these factors is calculated using Eq. 5. The LSI of each pixel within the study area was calculated by summing the contrast values of all conditionally independent landslide predictive factors (Eq. 19). After that, the LSI was normalized between 0 and 1. Finally, The LS map was divided into 5 susceptibility levels (namely, “very high”, “high”, “moderate”, “low” and “very low”) using the “natural breaks” classification method in ArcGIS ver10.2. As shown in Fig. 6a, 25.60% and 45.17% of the study area show “very high” and “high” proneness to landsliding, respectively. The “low” and “very low” areas covered only 9.85% and 2.32% of the total area, respectively.

$$\mathrm{LSI}_{\mathrm{WoE}}^{(x,y)} = C_{\mathrm{Elevation}}^{(x,y)} + C_{\mathrm{Slope}}^{(x,y)} + C_{\mathrm{DtoR}}^{(x,y)} + C_{\mathrm{PGA}}^{(x,y)} + C_{\mathrm{SPI}}^{(x,y)} \quad (19)$$

where C is the contrast weights of a given pixel spatially located at (x, y).

Results of the LR

As illustrated in Table 8, the −2-log likelihood decreases from 33,271.404 at the first step to 29,628.017 at the final step. Nine factors were included in the LR model, including aspect, elevation, slope, curvature, distance to the river, PGA, SPI, PCT and TWI. The lithology factor was not significantly related to the landslides at the given confidence level (95%). Cox and Snell’s R^2 was observed ranging from 0.264 to 0.648 and Nagelkerke’s R^2 ranged from 0.311 to 0.719. A high value of these two indexes means high accuracy of the model. Table 9 shows the coefficients of all factors finally kept in the regression. Coefficients of aspect, elevation, slope, PGA and TWI are positive, indicating these factors were positively related to the landslide occurrences. The highest coefficient of 6.011(positive) was assigned to PGA, while the

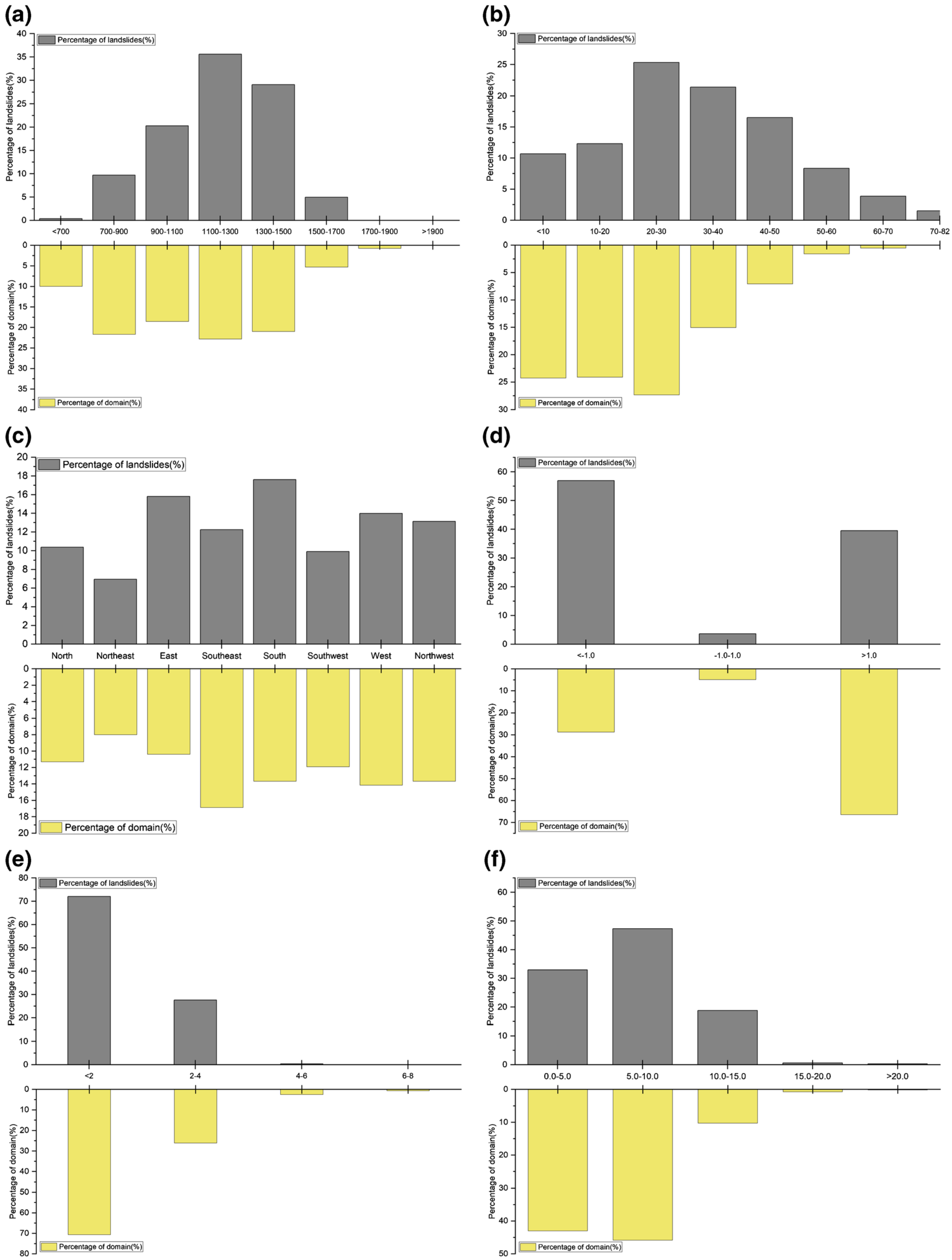

Fig. 5 Correlations between predictive factors and landslide distribution: **a** elevation; **b** slope gradient; **c** aspect; **d** curvature; **e** distance to the river; **f** TWI; **g** SPI; **h** lithology; **i** percentage of tree cover; and **j** PGA

lowest one (negative) was detected for the factor distance to the river. Both the factors TWI and SPI had little impact on landsliding, having coefficients of 0.112 and −0.103, respectively.

Ultimately, by assigning the coefficients to the variables used in the LR model, a weighted linear combination of landslide predictive factors was carried out in ArcGIS using Eq. 20. In order to create the landslide susceptibility map, the LSI

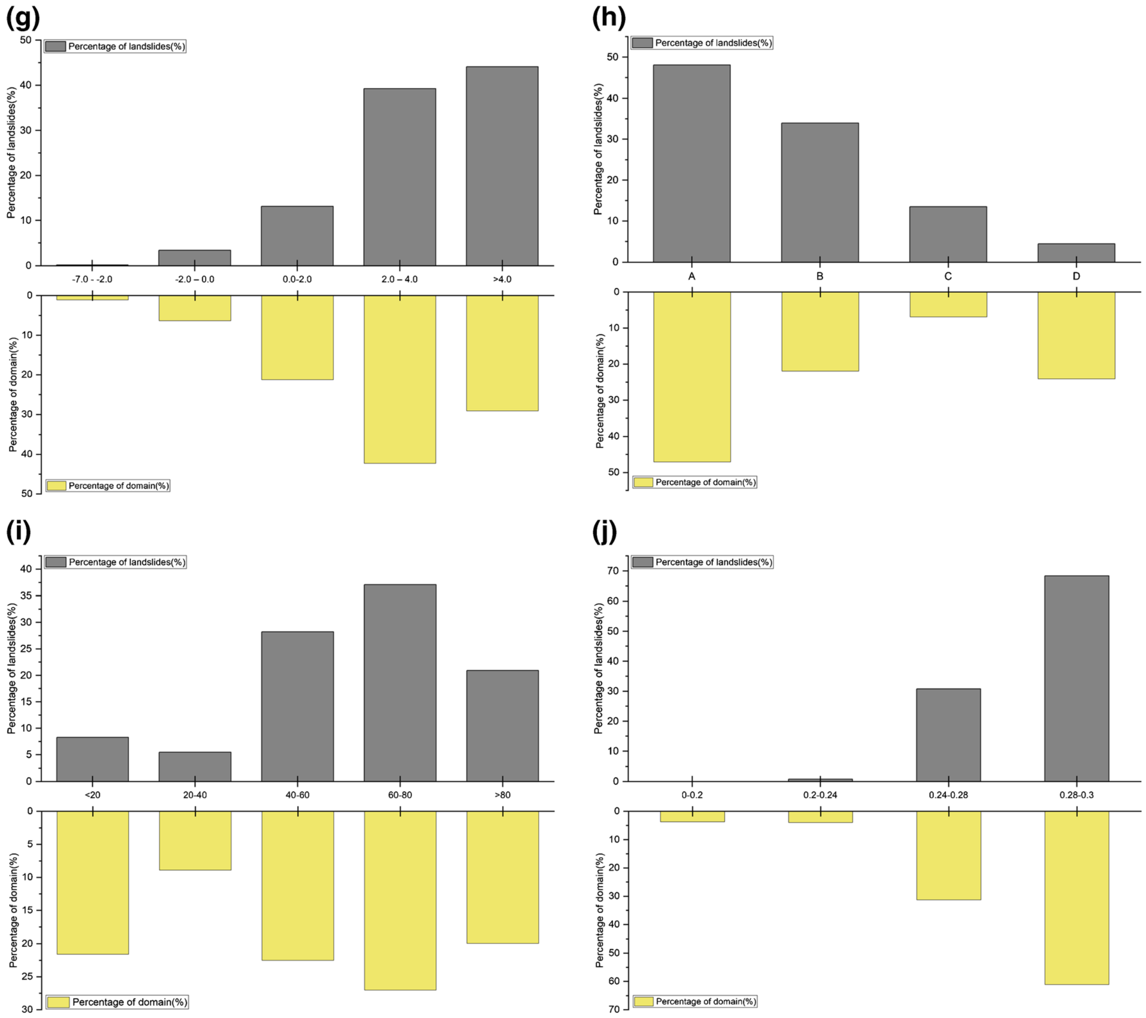

Fig. 5 (continued)

for each pixel with the study area was calculated by inserting Eq. 20 into Eq. 7. The obtained map was categorized into five susceptibility zones using the "natural breaks" methods. As indicated in Fig. 6b, the "moderate" susceptibility category took the biggest portion (28.80%) of the study area. 10.56% and 25.53% of the total area had "very high" and "high" susceptibilities of landsliding, respectively. Meanwhile, 19.48% and 15.62% of the total area were mapped as "very low" and "low" landside-susceptible regions, respectively.

$$p = \beta_{\mathrm{Aspect}}{}^{*}\mathrm{Aspect} + \beta_{\mathrm{Elevation}}{}^{*}\mathrm{Elevation} + \beta_{\mathrm{Slope}}{}^{*}\mathrm{Slope} + \beta_{\mathrm{PGA}}{}^{*}\mathrm{PGA} + \beta_{\mathrm{TWI}}{}^{*}\mathrm{TWI} + \beta_{\mathrm{DtoR}}{}^{*}\mathrm{DtoR} + \beta_{\mathrm{PTC}}{}^{*}\mathrm{PTC} + \beta_{\mathrm{SPI}}{}^{*}\mathrm{SPI} + \mathrm{Constant} \qquad (20)$$

where β is the regression coefficient in LR.

Results of the SVM

Through the aforementioned grid searching and MFCV, the highest accuracy 89.3% was obtained with C = 8 and γ = 0.125, respectively. The final LS map was created by entering each pixel with 10 attributes into the SVM model using RBF kernels with optimized C and γ values. The output value of SVM for each pixel ranged between 0 (indicating 0% likelihood of landsliding) and 1 (indicating 100% likelihood of landsliding). The final LS map using SVM was also divided into five susceptible classes using the "natural breaks" methods. As illustrated in Fig.6c, 7.94% and 11.48% of the study area belong to the "very high" and "high" category, respectively. The coverage percentages of "moderate", "low" and "very low" susceptibility category were 32.94%, 32.23% and 15.41%, respectively.

Accuracy validation

The ROC results are illustrated in Fig. 7; the LR model provides the highest AUC values for the training dataset (0.782) and the validation dataset (0.764). For the training dataset, the lowest AUC of 0.733 was observed for the WoE model. In the

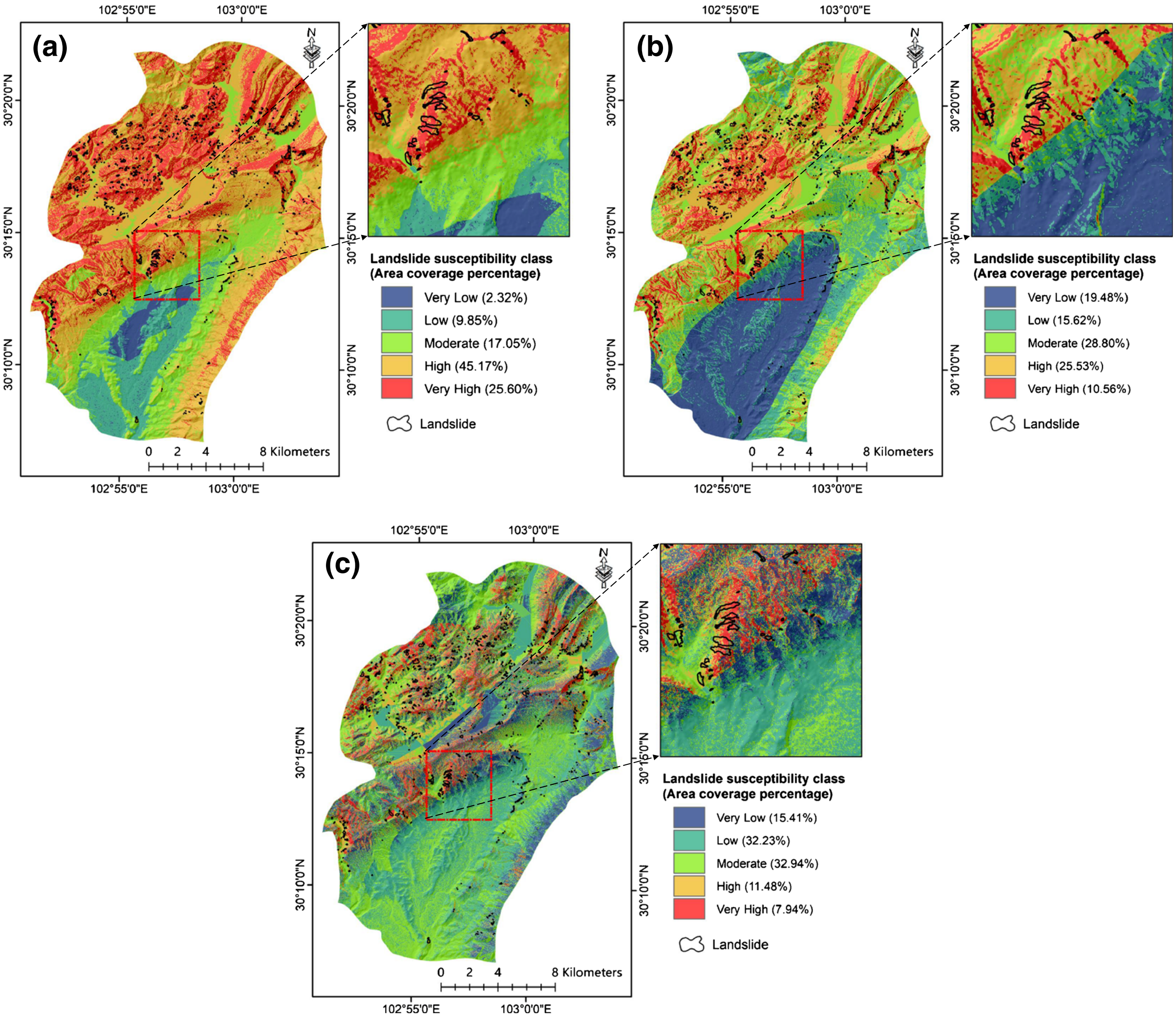

Fig. 6 Landslide susceptibility maps obtained using **a** WoE, **b** LR and **c** SVM

Table 8 The −2-log likelihood, Cox and Snell's and Nagelkerke's R^2 estimators for the LR

Step	−2 log likelihood	Cox and Snell R^2	Nagelkerke's R^2
1	33,271.404	0.264	0.311
2	31,387.148	0.316	0.347
3	30,738.531	0.388	0.402
4	30,325.984	0.427	0.451
5	29,938.826	0.469	0.495
6	29,807.918	0.512	0.538
7	29,702.657	0.548	0.591
8	29,638.83	0.563	0.624
9	29,628.017	0.648	0.719

Table 9 The coefficient and significance of each factor in the LR model

Factors	Coefficient	Significance
Aspect	0.502	0.000
Elevation	1.053	0.000
Slope	2.636	0.000
Curvature	--	0.780
DtoR	−1.648	0.000
Lithology	--	0.610
PGA	6.011	0.000
SPI	−0.103	0.000
PTC	−0.402	0.000
TWI	0.112	0.000
Constant	−6.935	--

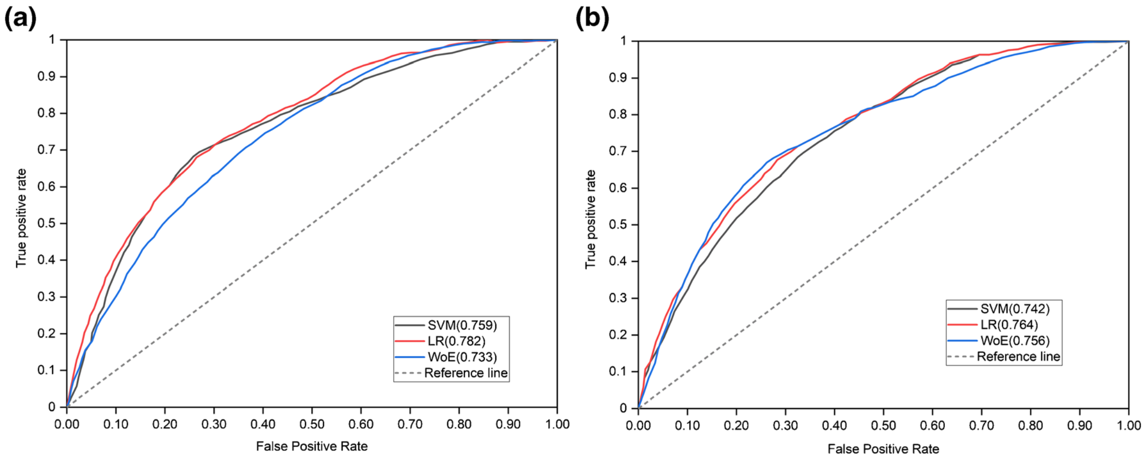

Fig. 7 Validation results of the models using ROC curves: (**a**) Training dataset; (**b**) Test dataset

case of the testing dataset, the SVM provided the lowest AUC of 0.742.

In order to compare the obtained LS maps produced using different models, the training and validation landslide datasets were overlaid with the LS maps, respectively. As illustrated in Fig. 8a, for all models, the landslide concentrations increased as the susceptibility level became high. For the training datasets, the result of the WoE shows that 58.03% and 38.41% of the landslide pixels occurred in the "very high" and "high" susceptibility category, respectively. As for the SVM model, the highest concentration (30.04%) of landslide pixels was found in the "high" category. In the case of LR model, the landslides percentages were 28.61%, 30.41%, 19.47%, 11.56% and 9.67% from "very high" level to "very low" level, respectively. Figure 8 b shows validation result using the testing dataset, and gradual increase of landslide concentration from "very low" to "very high" is observed for the LR model, with the highest percentage of 39.58% in "very high" and the lowest percentage of 1.79% in "very low". In terms of WoE, 46.07% and 43.25% of the total landslide pixels fell within the "very high" and "high" susceptible regions, respectively. As for the SVM, the "high" category had the most landslide pixels (31.41%), followed by the "very high" category (28.1%). However, there were 6.45% of the total landslide pixels observed in the "very low" category using the SVM.

Discussions

Validation using both training dataset and validation dataset showed that no obvious differences were observed among the accuracies of the three models. If validated by training dataset, the accuracy varied from 73.3% (the WoE) to 78.2% (the LR), while it varied from 74.2% (the SVM) to 76.4% (the LR) for the testing dataset. Despite the fact that more than 90% of the landslides fell within the "very high" (58.03%) and "high" (38.41%) susceptibility zones mapped using the WoE, an overestimation of LSI might have occurred, since the resulting WoE map may fail to facilitate the LSM for this study as sum of the "very high" and "high" landslide-susceptible regions had exceeded 70% of the total area. A possible reason

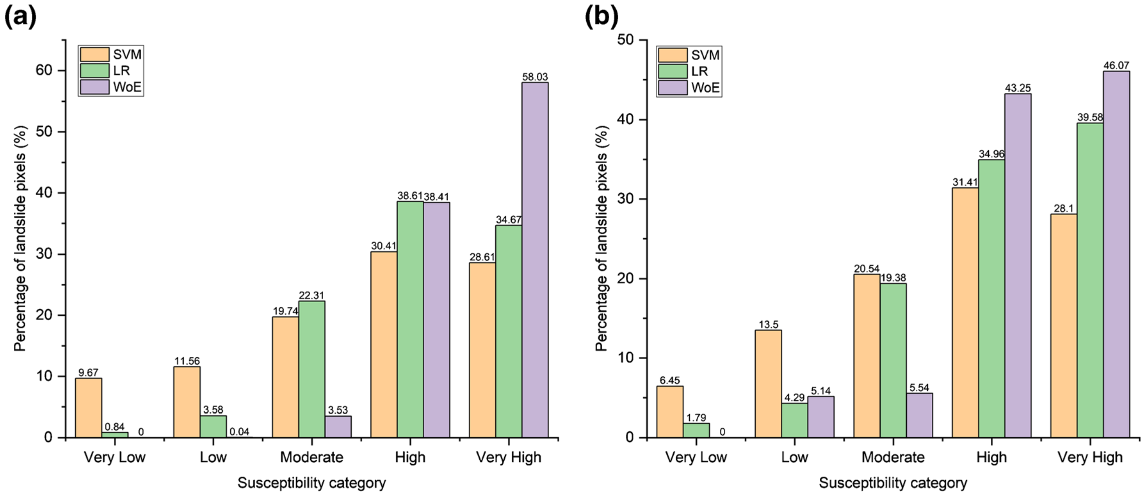

Fig. 8 Distribution of landslides in each landslide susceptibility category. **a** Training dataset; **b** validation dataset

accounts for such conditions may arise from the ignorance of the nonlinear correlations between landslides and their predictive factors. A predictive factor with a big contrast difference of different classes would significantly affect the final WoE wights, since all predictive factors were treated equally. Despite no conditional independence or specific distribution pattern (e.g., normal distribution) of the input data was required for the SVM, there is still room to improve the SVM through optimization of its kernels and the corresponding parameters (Kavzoglu et al. 2015). In addition, the implementation of the SVM in GIS modelling remains a challenging issue (Pham et al. 2016). It is noteworthy that several ensemble models have proposed and succeeded in most recent literature for LSM in order to utilize the advantages of individual models (Wang et al. 2014; Arabameri et al. 2019a, b; Lee 2019). Based on a general consideration of the given conditions discussed above, the LR and the SVM are more preferably recommended for LSM of the study area than the WoE.

In recent years, many studies have examined the efficiency of various machine-learning algorithms in LSM (Kalantar et al. 2017; Nhu et al. 2020; Yu and Chen 2020). The neural network (NN) and SVM were the two most frequently used algorithms, which were firstly developed to solve nonlinear classification problems. Nevertheless, it is more difficult to distinguish the good from the bad than to distinguish the better from the good. For example, Kalantar et al. (2017) and Yu and Chen (2020) suggested that SVM outperformed than ANN, while Işık (2010) and Nhu et al. (2020) found ANN showed a higher accuracy than SVM. Hence, a comparative study between LSM models for each case would help to optimize the result. Moreover, hybrid use of different algorithms has become increasingly popular in more recent years in making use of the advantage of each individual method. Such kind of hybrid methods had shown outstanding performance than individual methods (Yan et al. 2019; Sahana et al. 2020).

Conclusions

The LSM over a broad area has proved a cost-effective way to mitigate landslide hazards. However, the efficiency and accuracy of different models should be compared and evaluated in order to obtain a reliable LS map of a given study area. In this study, performances of three types of statistical LSM models were evaluated, including the bivariate WoE, the multivariate LR and a machine-learning model SVM, in seismic-prone regions of Sichuan Province, China.

An historical inventory of earthquake-induced landslides was utilized and randomly divided into training (70%) and testing (30%) groups. Ten factors were selected as an input of LSM models to develop landslide susceptibility maps. All data were elaborated in a GIS environment. In the obtained final LS maps, the study area had been classified into five different susceptibility zones. Although the LR model provided the highest accuracy among the three models, validation results show that there are no obvious differences in the accuracy of the obtained maps. In general, all the models had shown good performance with AUC values greatly than 0.75, when verified using the validation dataset, while all the AUC values were greater than 0.70, when verified using the training dataset.

Possible overestimation was detected in the maps produced using the WoE. Therefore, the LR and the SVM are found more suitable for the LSM in the study area. The resultant maps of this study can be of great help for the decision-makers in urban planning, for hazard mitigations of landslides and for the identification of existing infrastructures potentially exposed to future landslide risk. Moreover, a precise comparative examining and suggestion of different kinds of statistical models carried out in this study will aid future model selection of LSM in similar terrain conditions for the purpose of landslide disaster risk reduction.

Abbreviations LSM, Landslide susceptibility mapping; LFZ, Longmenshan fault zone; SDF, Shuangshi-Dachuan fault; GIS, Geographic information system; DEM, Digital elevation model; TWI, Topographic wetness index; SPI, Stream power index; PGA, Peak ground acceleration; PGA, Peak ground acceleration; PTC, Percentage of tree cover; WoE, Weight of evidence; LR, Logistic regression; SVM, Support vector machine; ROC, Receiver operating characteristic; AUC, Areas under ROC curves; FR, Frequency ratio; IV, Information value; RBF, Radial basis function; CNN, Convolutional neural network; CAS, Chinese Academy of Science; LSI, Landslide susceptibility index; TP, True positive; TN, True negative; FP, False positive; FN, False negative; FPR, False positive rate; TPR, True positive rate; TOL, Tolerances; VIF, Variance inflation variables; MLR, Maximum likelihood ratio; LN, Linear; PL, Polynomial; SIG, Sigmoid

Acknowledgements All financial supports were greatly acknowledged.

Funding This research was supported by the National Natural Science Foundation of China (51708199); Science and Technology Infrastructure Program of Guizhou Province (2020-4Y047; 2018-133-042); Fundamental Research Funds for the Central Universities (531118010069); Science and Technology Project of Transportation and Communication Ministry of Guizhou Province (2017-143-054); and Science and Technology Program of Beijing (Z181100003918005).

Declarations

Conflict of interest The authors declare that they have no competing interests.

References

Arabameri A, Pradhan B, Rezaei K et al (2019a) GIS-based landslide susceptibility mapping using numerical risk factor bivariate model and its ensemble with linear multivariate regression and boosted regression tree algorithms. J Mt Sci 16:595–618. https://doi.org/10.1007/s12665-018-7704-z

Arabameri A, Pradhan B, Rezaei K, Conoscenti C (2019b) Gully erosion susceptibility mapping using GIS-based multi-criteria decision analysis techniques. Catena 180:282–297. https://doi.org/10.1016/j.catena.2019.04.032

Bonham-Carter GF (1994) Geographic information systems for geoscientists. Pergamon Press, Oxford

Chapi K, Singh VP, Shirzadi A, Shahabi H, Bui DT, Pham BT, Khosravi K (2017) A novel hybrid artificial intelligence approach for flood susceptibility assessment. Environ Model Softw 95:229–245. https://doi.org/10.1016/j.envsoft.2017.06.012

Chen Y, Yang Z, Zhang Y et al (2013) From 2008 Wenchuan earthquake to 2013 Lushan earthquake. Sci Sin Terrae 43(6):1064–1072. https://doi.org/10.1007/s11430-013-4642-1

Chen W, Xie X, Peng J, Wang J, Duan Z, Hong H (2017) Gis-based landslide susceptibility modelling: a comparative assessment of Kernel logistic regression, Nave-Bayes tree, and alternating decision tree models. Geomat Nat Haz Risk 8(2):950–973. https://doi.org/10.1080/19475705.2017.1289250

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3): 273–297

Fan X, Juang CH, Wasowski J, Huang R, Xu Q, Scaringi G, van Westen CJ, Havenith HB (2018) What we have learned from the 2008 Wenchuan Earthquake and its aftermath: a decade of research and challenges. Eng Geol 241:25–32. https://doi.org/10.1016/j.enggeo.2018.05.004

Fanos AM, Pradhan B, Alamri A, Lee CW (2020) Machine learning-based and 3D kinematic models for rockfall hazard assessment using LiDAR data and GIS. Remote Sens 12(11):1755. https://doi.org/10.3390/rs12111755

Hong H, Tsangaratos P, Ilia I, Liu J, Zhu AX, Chen W (2018) Application of fuzzy weight of evidence and data mining techniques in construction of flood susceptibility map of Poyang County, China. Sci Total Environ 625:575–588. https://doi.org/10.1016/j.scitotenv.2017.12.256

Huang Y, Zhao L (2018) Review on landslide susceptibility mapping using support vector machines. Catena 165:520–529. https://doi.org/10.1016/j.catena.2018.03.003

Işık Y (2010) Comparison of landslide susceptibility mapping methodologies for Koyulhisar, Turkey: conditional probability, logistic regression, artificial neural networks, and support vector machine. Environ Earth Sci 61(4):821–836

Javdanian H, Pradhan B (2019) Assessment of earthquake-induced slope deformation of earth dams using soft computing techniques. Landslides 16(1):91–103. https://doi.org/10.1007/s10346-018-1078-x

Jena R, Pradhan B, Beydoun G, Alamri AM, Ardiansyah N, Sofyan H (2020) Earthquake hazard and risk assessment using machine learning approaches at Palu, Indonesia. Sci Total Environ 749:141582. https://doi.org/10.1016/j.scitotenv.2020.141582

Kalantar B, Pradhan BT, Naghibi SA, Motevalli A, Mansor S (2017) Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (svm), logistic regression (lr) and artificial neural networks (ann). Geomat Nat Haz Risk 9:1–21

Kavzoglu T, Kutlug Sahin E, Colkesen I (2015) An assessment of multivariate and bivariate approaches in landslide susceptibility mapping: a case study of Duzkoy district. Nat Hazards 76(1):471–496. https://doi.org/10.1007/s11069-014-1506-8

Khosravi K, Pham BT, Chapi K, Shirzadi A, Shahabi H, Revhaug I, Prakash I, Tien Bui D (2018) A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. Sci Total Environ 627:744–755. https://doi.org/10.1016/j.scitotenv.2018.01.266

Kumar AS, Kumar A, Krishnan R et al (2017) Soft computing in remote sensing applications. Proc Nat Acad Sci India Sect A Phys Sci 87(4): 503–517. https://doi.org/10.1007/s40010-017-0431-0

Lee S (2019) Current and future status of GIS-based landslide susceptibility mapping: a literature review. Korean J Remote Sens 35(1): 179–193. https://doi.org/10.7780/kjrs.2019.35.1.12

Nhu VH, Shirzadi A, Shahabi H, Singh SK, Al-Ansari N, Clague JJ et al (2020) Shallow landslide susceptibility mapping: a comparison between logistic model tree, logistic regression, Nave Bayes tree, artificial neural network, and support vector machine algorithms. Int J Environ Res Public Health 17(8):2749

Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830. https://doi.org/10.1524/auto.2011.0951

Pham BT, Pradhan B, Bui DT et al (2016) A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India). Environ Model Softw 84:240–250. https://doi.org/10.1016/j.envsoft.2016.07.005

Pham BT, Prakash I, Khosravi K et al (2018) A comparison of support vector machines and Bayesian algorithms for landslide susceptibility modelling. Geocarto Int 34(13):1385–1407. https://doi.org/10.1080/10106049.2018.1489422

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

Sahana, M, Pham, BT, Shukla, M, Costache, R. D., & Prakash, I. (2020). Rainfall induced landslide susceptibility mapping using novel hybrid soft computing methods based on multi-layer perceptron neural network classifier. Geocarto International (3).

Schicker R, Moon V (2012) Comparison of bivariate and multivariate statistical approaches in landslide susceptibility mapping at a regional scale. Geomorphology 161:40–57. https://doi.org/10.1016/j.geomorph.2012.03.036

Shahabi H, Ahmad BB, Khezri S (2013) Evaluation and comparison of bivariate and multivariate statistical methods for landslide susceptibility mapping (case study: Zab basin). Arab J Geosci 6:3885–3907. https://doi.org/10.1007/s12517-012-0650-2

Shrestha S, Kang T-S, Suwal MK (2017) An Ensemble Model for Co-seismic landslide susceptibility using GIS and random forest method. ISPRS Int J Geoinf 6(11):365. https://doi.org/10.3390/ijgi6110365

Sorensen R, Zinko U, Seibert J (2006) On the calculation of the topographic wetness index: evaluation of different methods based on field observations. Hydrol Earth Syst Sci 10:101–112. https://doi.org/10.5194/hess-10-101-2006

Sujatha ER, Rajamanickam GV (2015) Landslide hazard and risk mapping using the weighted linear combination model applied to the Tevankarai stream watershed, Kodaikkanal, India. Hum Ecol Risk Assess 21(6):1445–1461. https://doi.org/10.1080/10807039.2014.920222

Tehrany MS, Pradhan B, Mansor S, Ahmad N (2015) Flood susceptibility assessment using GIS-based support vector machine model with different kernel types. Catena 125:91–101. https://doi.org/10.1016/j.catena.2014.10.017

Wang M, Liu M, Yang S, Shi P (2014) Incorporating triggering and environmental factors in the analysis of earthquake-induced landslide hazards. Int J Disaster Risk Sci 5(2):125–135. https://doi.org/10.1007/s13753-014-0020-7

Yan F, Zhang Q, Ye S, Ren B (2019) A novel hybrid approach for landslide susceptibility mapping integrating analytical hierarchy process and normalized frequency ratio methods with the cloud model. Geomorphology 327(FEB.15):170–187

Yu C, Chen J (2020) Landslide susceptibility mapping using the slope unit for Southeastern Helong City, Jilin Province, China: A Comparison of ANN and SVM. Symmetry 12:1047. https://doi.org/10.3390/sym12061047