ORIGINAL PAPER

The performance quality of LR, SVM, and RF for earthquake-induced landslides susceptibility mapping incorporating remote sensing imagery

Rui Liu[1,2] · Luyao Li[1] · Saied Pirasteh[3] · Zili Lai[4] · Xin Yang[1] · Himan Shahabi[5,6]

Abstract

An earthquake with Ms 7.0 (33.2° N, 103.8° E) occurred in Jiuzhaigou County of Sichuan Province in China on 8 August 2017. This earthquake triggered a large number of landslides in the study area. Although the susceptibility quality level index has improved, the high-quality assessments still have remained rare. We adopted three models, including the logistic regression (LR), support vector machine (SVM), and random forest (RF) to study the quality performance of the susceptibility distribution rule of earthquakes induced landslides. We used satellite images of before and after earthquakes and landslides as well. We used the area under receiver operating characteristic (ROC) curve (AUC) and ratio to evaluate the model's accuracy and quality performance, including the mapping availability susceptibility assessment. This study reveals that RF has the highest ratio (2.07) as compared to the LR (1.78) and SVM (1.90). The result shows that RF has more potential to implement future experiments in Sichuan Province because of a better performance quality in the susceptibility assessment of landslides induced by earthquakes.

Keywords Earthquake-induced landslide · Random forest (RF) · Logistic regression (LR) · Support vector machine (SVM) · ROC curve

Introduction

Sometimes, landslides become critical natural issues in the Sichuan plateau, mostly the giant panda sanctuaries. Some of the landslides in Sichuan Province are secondary disasters that are triggered by earthquakes. They generate casualties and damage to infrastructure and environmental, historical, and cultural heritage sites (Cao et al. 2019). Worse still, there would be a severe implication for the habitat for the wild giant pandas. On 8 August 2017, a Ms 7.0 earthquake occurred in Jiuzhaigou County, Aba Autonomous Prefecture of Sichuan Province, China

✉ Saied Pirasteh
sapirasteh@swjtu.edu.cn

Rui Liu
lr@cdut.edu.cn

Luyao Li
leeleoo@163.com

Zili Lai
1049962421@qq.com

Xin Yang
kkshiney@outlook.com

Himan Shahabi
h.shahabi@uok.ac.ir

[1] College of Geophysics, Chengdu University of Technology, Sichuan 610059, China

[2] Key Lab of Earth Exploration and Information Techniques of Ministry Education, Chengdu University of Technology, Sichuan 610059, China

[3] Department of Surveying and Geoinformatics, Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, The Western Park of the Hi-Tech Industrial Development Zone, Chengdu 611756, Sichuan, China

[4] Dongpo District, Meishan, Sichuan Province, Housing and Urban-Rural Construction Bureau, Meishan 620000, Sichuan, China

[5] Department of Geomorphology, Faculty of Natural Resources, University of Kurdistan, Sanandaj, Iran

[6] Department of Zrebar Lake Environmental Research, Kurdistan Studies Institute, University of Kurdistan, Sanandaj, Iran

(location: 33.2° N, 103.8° E) and focal depth of 20 km (CENC 2017).

In the early 1960s, many countries have realized the importance of earthquake landslides susceptibility mapping, and enormous research has been published around the world (Carrara 1983; Brabb et al. 1972; Reichenbach et al. 2018). Many researchers have tried to evaluate landslides susceptibility utilizing different types of methods; for example, some of them are qualitative analysis, semi-quantitative analysis, and quantitative analysis (Lee et al. 2008; Wang and Lin 2010; Lee and Talib 2005; Li et al. 2017a, b; Reichenbach et al. 2018; Pirasteh et al. 2020; Zhu et al. 2020). These studies analyzed the relationships between landslides and natural predisposing factors. They also evaluated landslide susceptibility and mapped the landslide susceptibility distribution. For example, Reichenbach et al. (2018) implemented a susceptibility quality level index. They found that though the quality of published models has been improved over the years; however, high-quality assessments still remained rare. Wang and Lin (2010) applied the infinite slope theory to the regional analysis and development of potential shallow landslide maps. Furthermore, Torgoev and Havenith (2013) used seismic slope stability and Newmark displacement method in 2D dynamic modeling to a landslide-prone area in the Mailuu-Suu Valley. Li et al. (2017a) showed the progressive failure of slope truly by calculating the warning deformation of landslides. Faris and Wang (2014) sought to examine the initiation mechanism of landslides that are triggered by earthquake during rainfall. They developed a new approach to evaluating landslide susceptibility through pore pressure strain and stiffness.

Liu et al. (2020) introduced the key technologies and machine learning algorithms in data collection, data storage, and data processing and summarized the applications of big data technology in landslides, mudslides, and other geological disasters by domestic and foreign scholars. An unsupervised representation learning module, which features independence, compactness, robustness, and transferability, was designed by Zhu et al. (2020). They restricted Boltzmann machines and denoising autoencoder to unsupervised discover the underlying representations embedded in the thematic maps. They applied the transferring strategy in an adversarial manner to generalize the learned representations to the sample-scarce area. Ye et al. (2019) studied landslides triggered by the earthquake and proposed a deep learning framework with constraints to detect landslides on the hyperspectral image. The framework consisted of two steps, including a deep belief network to extract the spectral-spatial features of a landslide, and then inserted the high-level features and constraints into a logistic regression classifier for verifying the landslide.

Besides, in recent years, many statistical methods, geographical information system (GIS), and remote sensing techniques have been used to analyze landslide susceptibility (Pirasteh et al.

2020; Reichenbach et al. 2018; Wang et al. 2019; Zhu et al. 2020), including the logistic regression (Ayalew and Yamagishi 2005; Erener and Düzgün 2012; Wei et al. 2013), frequency ratio (FR) (Juliev et al. 2019; Yan et al. 2019), and index of entropy (Jaafari et al. 2014; Shirani et al. 2018). Although some of the researchers have combined statistical approach and GIS techniques and performed more suitable for the large study area, landslide susceptibility mapping is a typical complex and nonlinear problem (Lai and Tsai 2019). However, the results from the statistical approach may not achieve satisfactory accuracy. Subsequently, many machine learning approaches are introduced, such as SVM (Pandit et al. 2017; Pawluszek et al. 2018; Wang et al. 2019; Hosseinalizadeh et al. 2019), random forest (Hong et al. 2016), artificial neural network (Lee et al. 2003; Pham et al. 2017), and decision trees (Chen et al. 2018). These methods have been successfully applied in many places for landslide susceptibility assessment incorporated with the GIS. Meten et al. (2015) applied FR and LR to landslides' susceptibility assessment by using the GIS. Zhou and Fang (2015) built SVM with four kernel function such as radial basis function, polynomial, linear, and sigmoid kernel function. They tested the accuracy of those models by generating landslide susceptibility maps and comparing them with known landslides (Zhou and Fang 2015). Dou et al. (2015) considered the spatial autocorrelation of landslide causative factors. They used CF to optimize the landslide causative factors.

The RF and LR models are two models that have been used commonly in many studies because they are effective in mapping landslide susceptibility (Wei et al. 2013). Currently, SVM is relatively mature in landslides susceptibility assessment (Wang et al. 2019). Nevertheless, these models' application to other areas is limited without testing the prediction accuracy and spatial generalization ability. Therefore, the performance quality and a clear geographical bias in susceptibility study locations remain unknown how well these models will perform in seismic landslides areas, respectively. To predict seismic landslide susceptibility in a new area by applying the models, it will make sense to learn about the natural disasters resulting from earthquakes. For this reason, we selected the RF, LR, and SVM models in this study. It is because the objective is to understand the potential application and performance quality of each model. Notably, we consider geology and environmental conditions incorporating remote sensing geological interpretation to develop the earthquake-triggered landslides susceptibility models of Jiuzhaigou.

Moreover, the application of machine learning methods in landslide disaster prevention is still not comprehensive in many scholars' current research, for example, selecting landslide influencing factors due to many factors that affect landslides. The selection of factors is often ignored or did not pay attention to the importance of analyzing the selected factors. On the other hand, the authors noticed that the hyperparameter optimization problem was ignored in the algorithm selection

process, and the cross-validation and model training was not emphasized. Similarly, in the final classification of landslide sensitivity, only one verification method is used to prove the mapping or model accuracy, leading to inevitable misjudgments.

To sum up, in addition to the high research value of selected areas, the research also made some remarkable improvements based on traditional machine learning methods for landslide susceptibility research: (1) used the variance-inflation-factor (VIF) to evaluate the collinearity of landslide-inducing factors and to confirm the high correlation; (2) evaluated the three kernel functions of support vector machines and selected the one which with the best performance, then compared with the other two algorithms to obtain more accurate and comprehensive comparison results; (3) used the root mean square error (RMSE) to select the parameters of the algorithm and using the best one to build the model to reduce the errors caused by the parameters; (4) we used two methods to evaluate the vulnerability of the algorithms. Therefore, we can comprehensively draw a comparison of the advantages and disadvantages of the models for the landslide zoning in the study area.

The geology of the study area is explained in the "Study area and geology" section. The "Data acquisition" section describes how the data are acquired. The processing of data and the methodology are explained in the "Method" section. This section explains the comparison of the methods and validation. The "Results" section provides the study results. We discuss how to use the techniques in this study, and which one has a better performance than another in the "Discussion" section. We conclude the paper in the "Conclusion" section.

Study area and geology

The study area locates in Jiuzhaigou County. It is in the northern boundary of Sichuan Province, and it situates in the east of the Bayan Har block on the eastern margin of the Qinghai–Tibetan Plateau (Fig. 1). The regional geology comprises of features from Devonian to Triassic outcrops, and it consists mostly of slate, limestone, dolomite, and metamorphic sandstone (Li et al. 2019). Geologically, the study area has relatively complex structures because of the tectonic activities and the stresses from different directions in the regional area (Wang et al. 2018). Morphologically, we found that the lithology, tectonic uplift, glaciation, and river erosion control structures in the study area. This geomorphology phenomenon results in considerable elevation differences and variable landforms. The terrain in the southern part of the region is relatively high. The northeastern terrain has low terrain ranging from 1788 to 4870 m above mean sea level. The valleys in the study area trend from SSW to NNE or NWW to SEE. They are canyons with steep slopes and high cliffs, making them

prone to landslides, mainly when it triggers by earthquakes. For example, an Ms 7.0 earthquake happened in Jiuzhaigou National Nature Reserve on August 8, 2017; it not only caused 25 deaths and property losses but it also destroyed paleontology fossils and ancient glacial landscapes. The earthquake's epicenter was at 33.20° N and 103.82° E with a 20-km depth (CENC 2017). It is considered the extension of the Huya fault (Li et al. 2017a, b; Zhang et al. 2018). This inferred seismogenic fault passes through the Fiver Flower Lake to Jiuzhai Paradise, shown in the NW–SE direction. A number of seismic geohazards, including landslides, debris flows, and rockfalls have been generated by the triggered-earthquake phenomenon. Moreover, it resulted in major injuries to the panda habitat, cultural heritage, and environment.

Data acquisition

We used remote sensing images, terrain, seismic, and regional planning data, in conjunction with some field observation sampling data. Remote sensing data include Ziyuan-3 surveying satellite image (ZY-3, 2.1 m) and Landsat-5 satellite image (30 m). We acquired the terrain data of Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model (GDEM) (V2.0, 30 m, UTM/WGS84) from NASA (http://www.nasa.gov/). The seismic data of the peak ground acceleration (PGA) is derived from the USGS (http://glovis.usgs.gov/). We collect the regional planning data from the State Bureau of Surveying and Mapping of Sichuan Province of China to define administrative divisions in the study area.

We used China satellite ZY-3 images (2.1 m, UTM/WGS84, spectral range: 0.50–0.80um, date: 03/31/2012, path number: 1248) as benchmarks to correct the drone data geometrically by using ENVI software. We interpreted the images of the Landsat 5 satellite (after the earthquake) based on the image interpretation techniques (Ali and Pirasteh 2004; Lillesand et al. 2015) and compare them with the ZY-3 satellite remote sensing image (before the earthquake) and field observations as well (Fig. 2). We prepared the datasets as different layers to compute them in ArcGIS and also to prepare the landslide susceptibility zonation map. The authors constructed a landslide inventory map with 270 landslides that are triggered by the earthquake (Fig. 3). We divided the landslide locations randomly into two parts, 70% were employed for the training model, and 30% applied to verify models.

Method

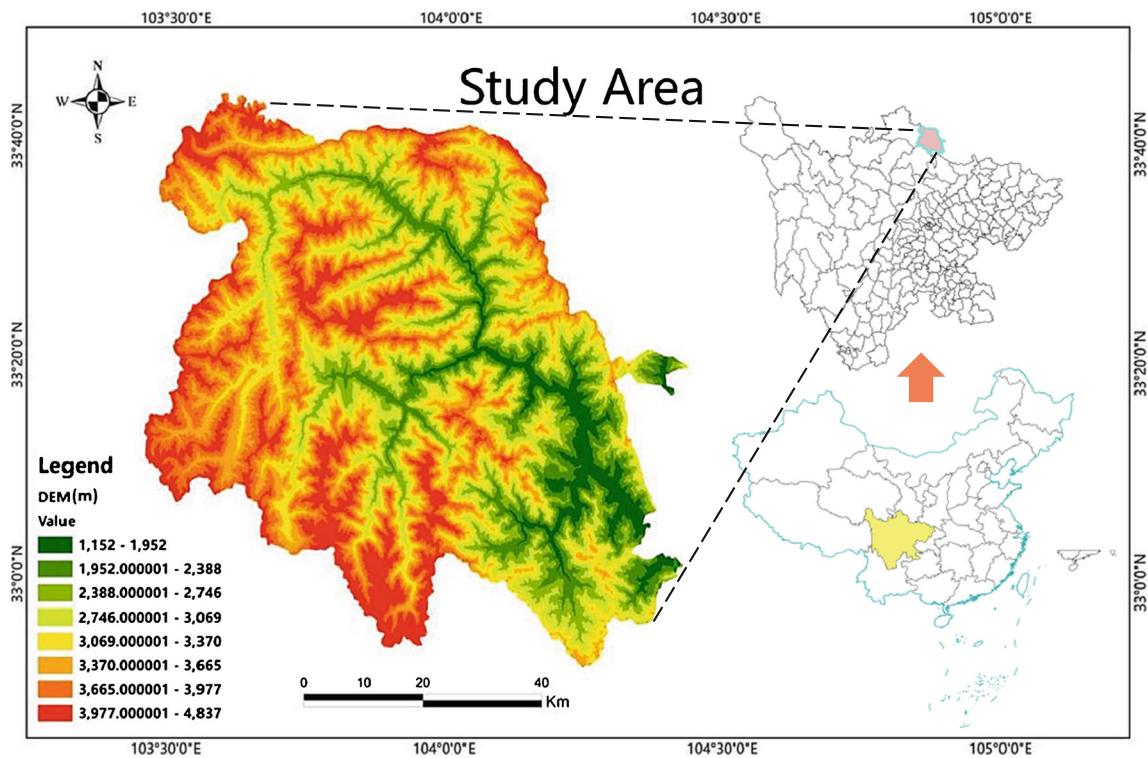The method contains four major sections. We created a geodatabase from the required data for landslide

Fig. 1 Study area

susceptibility mapping. The previously required maps, such as geology maps, were scanned and digitized on-screen by using ENVI and ArcGIS. The coordinate conversion process was applied to generate the same map coordinates. We used digital image processing techniques for atmospheric and geometric corrections of satellite and drone images (Lillesand et al. 2015). The image element techniques were applied to interpret the images and to extract the required features such as lineaments (Ali and Pirasteh 2004; Lillesand et al. 2015). In the next section, we constructed three models to develop a susceptibility map and compared the three models. Further**more**, in the last stage, we evaluated each model's performance quality (Fig. 4).

Landslide predisposing factors

The occurrence of landslides depends on many factors, and the relationship between these factors is complex (Reichenbach et al. 2018). According to Dai et al. (2002) and Wan (2009) landslide, predisposing factors are two categories: (1) the static predisposing factors include the environment of mountains that decides the potential sensitivity of landslides. (2) The dynamic predisposing factors include rainfall and earthquake. These dynamic predisposing factors may trigger landslides to some degree. Built on the above, we selected the following static predisposing factors. (a) The slope degree, (b) relief degree of the land surface, (c) normalized difference vegetation index (NDVI), and (d) land use

Fig. 2 Field photos from the study area (**a**) woodland area, (**b**) along the road

Fig. 3 Landsat image of study area
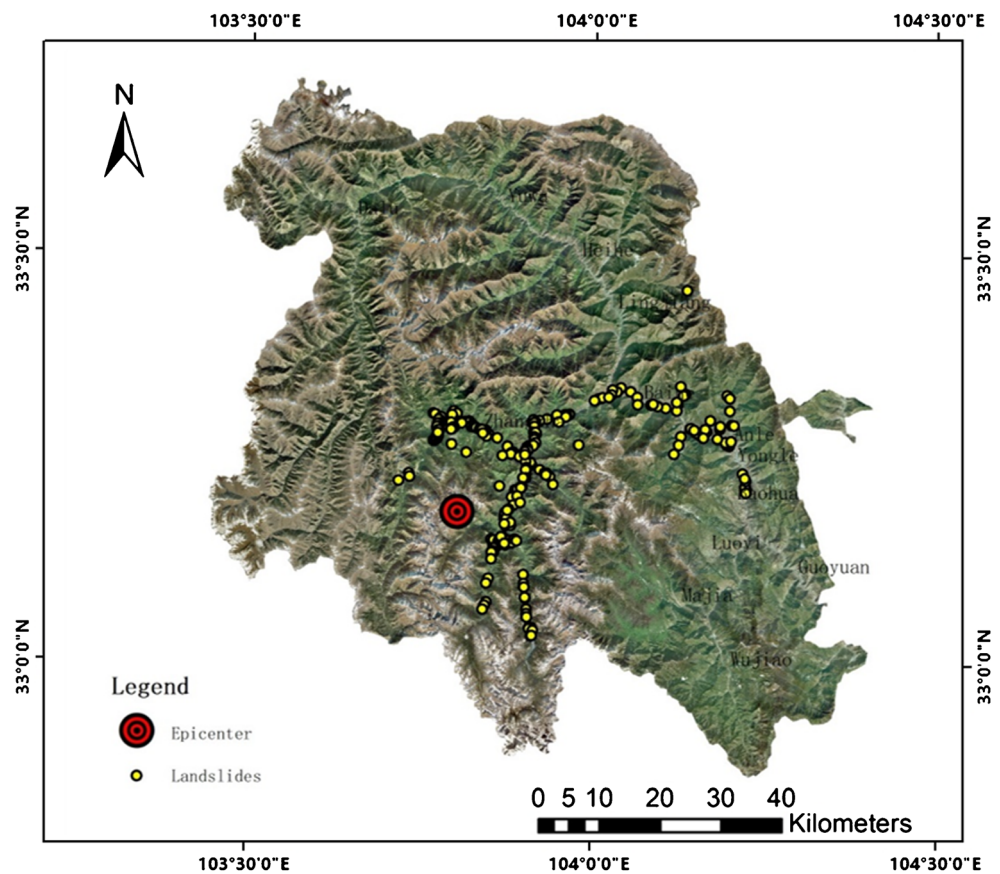
types, (e) peak ground acceleration (PGA), and (f) the distance to the seismic faults.

(a) The slope degree (SLOPE) is a dominant impact factor for landslides susceptibility (Wu et al. 2004; Kaya et al. 2015). According to the mechanical analysis of landslides, the mountain is the decline in the power of $G \times \sin$, where G is the gravitational potential energy,   is the terrain slope. In the same situation, the greater value of the slope degree triggers more landslides. We extracted the slope degree from the ASTER GDEM (V2.0, 30 m, UTM/WGS84) by ArcGIS based on the subset size of $3° \times 3°$and, the average maximum technique. The steep slope is mainly distributed in the northwest area of Lingguan town, and 73% of landslides have occurred on slopes over 20° (Fig. 5a).

(b) A relief degree of the land surface (RDL) can be an excellent performance characteristic as a whole. The relief degree of land surface values is expressed by the formula of $H_{max}-H_{min}$, where H_{max} is the highest altitude, and H_{min} is the lowest altitude (Feng et al. 2008). A landslide can be abstracted as inhomogeneous slope block material. Previous studies (Pirasteh et al. 2020) have shown that the landslide susceptibility tends to

increase with the relief degree of land surface values (Fig. 5b).

(c) Normalized difference vegetation index (NDVI) represents the fraction of the area's vegetation coverage (Lee and Talib 2005; Ahmed 2015). This study selects the Landsat-5 data to extract vegetation information in the study area before the earthquake (Fig. 5c). We applied atmospheric correction, geometric correction, and band operation digital image processing techniques (Lillesand et al. 2015). The NDVI computation formula is as follows:

$$NDVI = \frac{NIR-R}{NIR + R} \tag{1}$$

where NIR is near-infrared band, and R is the visible red band.

(d) The land use types (LT) are one of the most sensitive factors (Bourenane et al. 2014). We divided the land use types into construction land, cultivated land, forest land, and meadows (Fig. 5d).

Fig. 4 Methodology flowchart
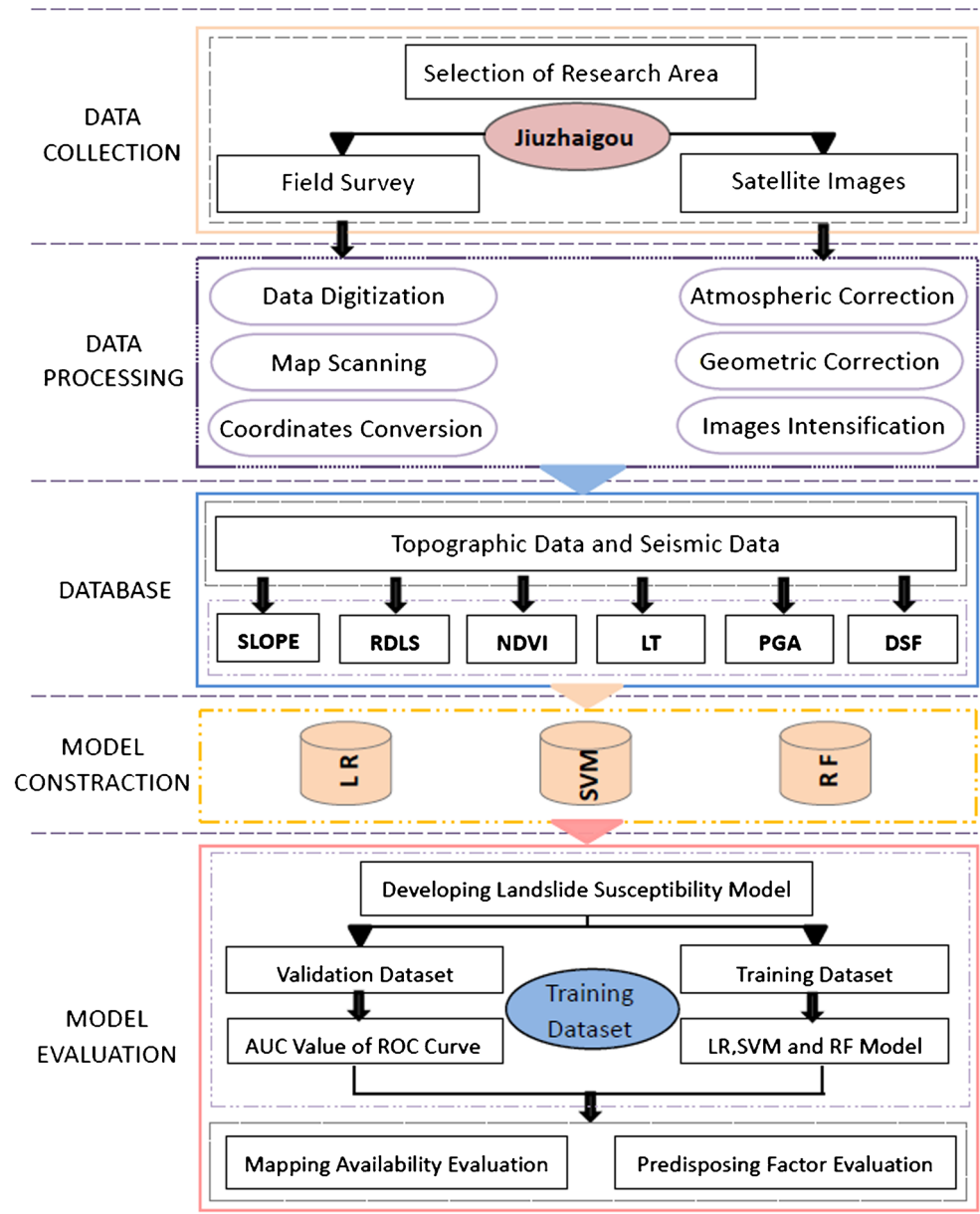
(e) Peak ground acceleration (PGA) is the acceleration of the ground movement. The greater value of the PGA causes greater stress to the mountain and decreases the stability of the mountains (Lee et al. 2008; Zhou and Fang 2015). We determined that 91% of the landslides have occurred in the PGA range larger than 0.28 g. Moreover, 56% of landslides took place in the range of PGA greater than 0.44 g (1 g = 10 m/s^2) (Fig. 5e).

(f) Distance from fault (DF) has a decisive role in the release of seismic energy. As we close to the fault, we will see more concentrated seismic energy (Farrokhnia et al. 2011; Xu et al. 2013). Generally, a lot of landslides disaster appears near the seismic fault (Elayaraja and Ganapathy 2015). We used the spatial analysis function

to calculate the distance to the seismic faults by using ArcGIS (Fig. 5f).

Predisposing factor classification

We employed the factor classification method to reduce the spatial autocorrelation of landslide predisposing factors. We divided the SLOPE, RDL, PGA, and NDVI into five classes (Table 1) based on the natural discontinuities classification method. DF and LT have been divided into four categories (Table 1).
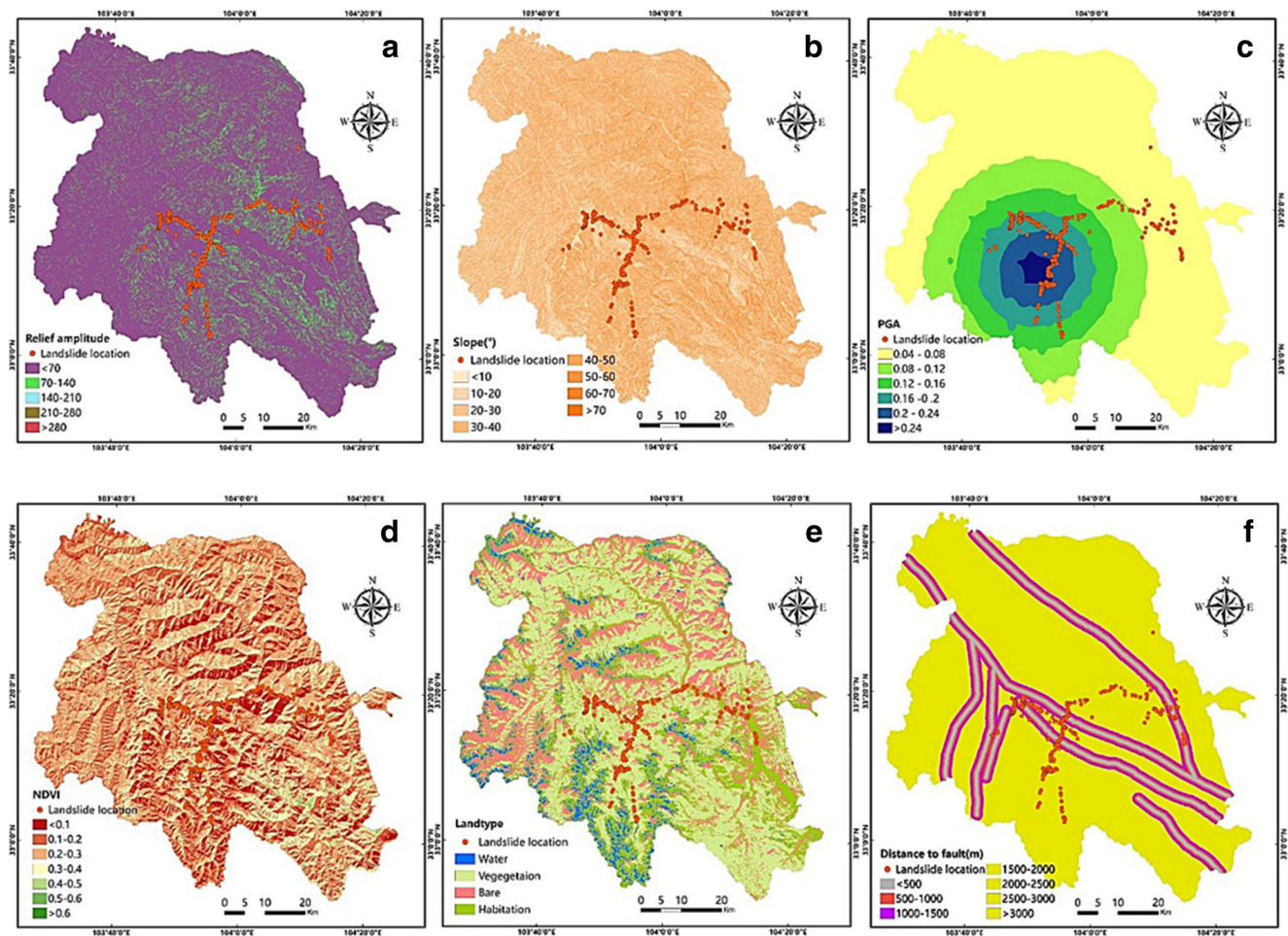
Fig. 5 Landslide predisposing factors. **a** Slope degree. **b** Relief degree of the land surface. **c** Peak ground acceleration (PGA). **d** Normalized difference vegetation index (NDVI). **e** Land use types. **f** The distance from the seismic fault

Table 1: Overview of classes limits for different factors

Model construction

We extracted 270 earthquake-induced landslides and 270 non-landslides (i.e., the susceptibility of landslides is about zero). We also extracted the predisposing factors. The FR method divided the total data into the training data (70%, 378 sample points) and test data (30%, 162 sample points). The training data were used to build the model, and the experimental data have used to verify the model's performance.

Logistic regression

The LR has been widely used in many areas such as biomedical, economic management, and geography (Yang et al. 2015). The logistic regression model is derived from the generalized linear model. This model combines response variables and independent variables by the connecting function.

Table 1 Overview of classes limits for different factors

Factor	Number of classes	Classes
SLOPE/(°)	5	1: (0–12), 2: (12–21), 3: (21–30), 4: (30–41), 5: > 41
RDL/(m)	5	1: (0–23), 2: (23–39), 3: (39–58), 4: (58–87), 5: > 87
PGA/(g)	5	1: (0–0.32), 2 : (0.32–0.40), 3: (0.40–0.48), 4: (0.48–0.52), 5: > 0.52
NDVI	5	1: < 0, 2 : > 0.6, 3 : (0.47–0.60), 4: (0.34–0.47), 5: (0–0.34)
LT	4	1: Building land, 2: arable land, 3: forest, 4: grass
DF/(m)	4	1: > 1200, 2 : (800–1200), 3: (400–800), 4: (0–400)

Table 2   Usage specification of the LibSVM and the operation process

Usage specification

- s svm_type	Set type of SVM (default 0): 0 – C-SVC 1 – nu-SVC 2 – one-class SVM 3 – epsilon-SVR 4 – nu-SVR
- t kernel_type	Set type of kernel function (default 2) 0 -- linear: u′*v 1 -- polynomial: (gamma*u′*v + coef0)^degree 2 -- radial basis function: exp(-gamma*|u-v|^2) 3 -- sigmoid: tanh(gamma*u′*v + coef0)
- d degree	Set degree in kernel function (default 3)
- g gamma	Set gamma in kernel function (default 1/num features)
- r coef 0	Set coef 0 in kernel function (default 0)
- c cost	Set the parameter C of C-SVC, epsilon-SVR, and nu-SVR (default 1)
- n nu	Set the parameter nu of nu-SVC, one-class SVM, and nu-SVR (default 0.5)
- p epsilon	Set the epsilon in loss function of epsilon-SVR (default 0.1)
- m cachesize	Set cache memory size in MB
- e epsilon	Set tolerance of termination criterion
- h shrinking	Whether to use the shrinking heuristics, 0 or 1
- b probability estimates	Whether to train a SVC or SVR model for probability estimates, 0 or 1
- wi weight	Set the parameter C of class i to weight*C, for C-SVC

We estimated the probability of a landslide by using the model output (it is between 0 and 1) for each grid cell. Therefore, we express the LR model as follows:

$$p(Y = 1) = \frac{1}{1 + e^{-z}} = \frac{e^{\alpha+\beta_1 x_1+\ldots+\beta_n x_n}}{1 + e^{\alpha+\beta_1 x_1+\ldots+\beta_n x_n}} \qquad (2)$$

The ratio between the probability p (i.e., dependent variable Y is 1) and the probability $(1 - p)$ (i.e., the dependent variable Y is 0) indicates the odds or likelihood ratio. The natural logarithm of odds (Logit) defines a linear expression of the explanatory variables $x_1 \ldots x_n$, and it is expressed by:

$$logit(p) = \ln\frac{p}{1-p} = \alpha + \sum_{j=1}^{n} \beta_j x_j$$
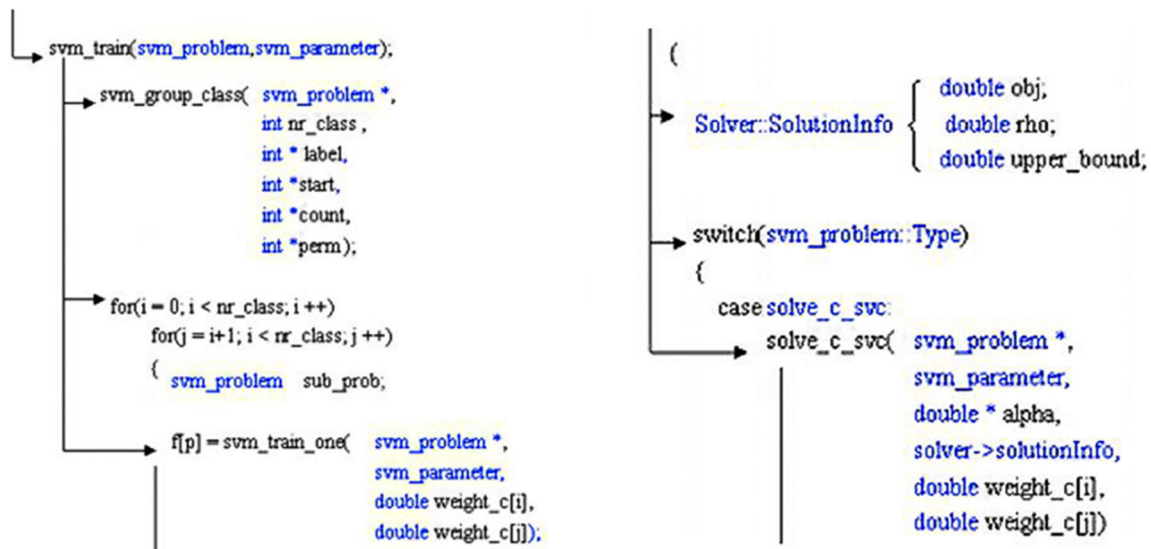$$= \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \qquad (3)$$

Fig. 6 Example of the algorithm for the operation process

Table 3 Confuse matrix

Confuse matrix	Actual value: P	Actual value: N
Prediction outcomes: P	TP	FP
Prediction outcomes: N	FN	TN

The intercept of the regression function is represented by α. The coefficient that measures the contribution of the independent variables (X_i, $i = 1 - n$) is $\beta i (i = 1 - n)$. We also used the maximum likelihood function to estimate the coefficient β. We consider that if the β is positive, the factor increases the probability of change. The factor has the opposite effect if β is negative. We utilized the statistical package for R to calculate all the coefficients.

Support vector machine

Vapnik (Vapnik and Cortes 1995) proposed SVM for regression in 1996. The SVM constructs a hyperplane or set of hyperplanes in a high or infinite-dimensional space. It can be used for classification and regression. A kernel function $K\left(\overrightarrow{x_i}, \overrightarrow{x_j}\right) = \varnothing\left(\overrightarrow{x_i}\right)^T \varnothing\left(\overrightarrow{x_j}\right)$ was used to account for the non-linear decision boundary (Vapnik and Cortes 1995). Here, training vectors $\overrightarrow{x_i}$'s are mapped into a higher (maybe infinite) dimensional space by function $\varnothing$. To solve a nonlinear problem, three kernel functions have been introduced commonly.

The following four types of kernel function are implemented in this study:

Linear kernel function (linear):

$$K\left(\overrightarrow{x_i}, \overrightarrow{x_j}\right) = \overrightarrow{x_i}^T * \overrightarrow{x_j} \tag{4}$$

Radial basis kernel function (RBF):

$$K\left(\overrightarrow{x_i}, \overrightarrow{x_j}\right) = \exp\left\{-/\delta^2 \left\|\overrightarrow{x_i} - \overrightarrow{x_j}\right\|^2\right\} \tag{5}$$

Table 4 Multicollinearity diagnostics

Factor	VIF	Tolerance
SLOPE	2.079183	0.4809582
RDL	2.047919	0.4883006
PGA	1.030851	0.9700727
NDVI	1.070813	0.9338698
LT	1.020979	0.9794522
DF	1.054336	0.9484646

where δ^2 is the bandwidth of the radial basis function.

Polynomial kernel function (polynomial):

$$K\left(\overrightarrow{x_i}, \overrightarrow{x_j}\right) = \left(\gamma X_i^T + r\right)^d \tag{6}$$

where the degree of the polynomial kernel and $\gamma > 0$ is d.

We derived and rasterized the environmental parameters by GIS techniques. This study employed the SVM analyses by using the LibSVM software. It provides a simple interface where analysts can easily link it with their own programs. The SVM model only outputs 0 and 1. The usage specification of the LibSVM and the algorithm operation process of calling the function are shown in Table 2 and Fig. 6.

Random forest

The RF has a series of decision trees. Each decision tree is independent, and it can get a result. The RF is the combination of these decision trees. In this way, the RF can achieve a better result than the decision tree model.

RF:

$$H(X) = \text{av}_k \max \sum_{i=1}^{k} I(h_i(X) = Y) \tag{7}$$

where, av_k indicates averaging and I is the indicator function.
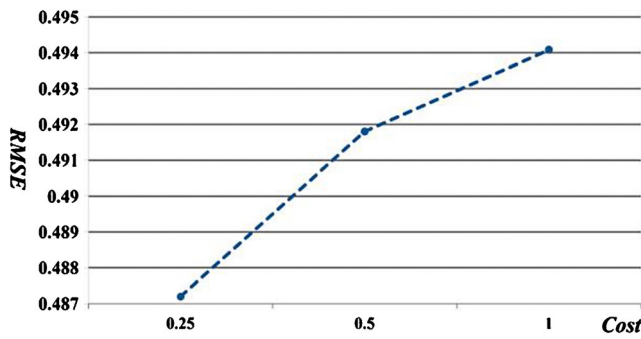
Fig. 7 The importance of landslide predisposing factors

Fig. 8 Parameter training of SVM (linear)

Margin function:

$$mg(X, Y) = av_k I\big((h_k(x) = Y) - \max_{j \neq i} av_k I(h_k(x) = j)\big) \quad (8)$$

The margin function is used to express the reliability of the model. The generalization error is given:

$$PE = P_{X,Y}(mg(x, y) < 0) \quad (9)$$

When increasing the decision tree, RF almost converges to:

$$PE = P_{XY}\big(av_k I(h_k(X) = Y) - \max_{j \neq i} av_k I(h_k(X) = j) < 0\big) \quad (10)$$

This result explains that the random forest with only one tree will over-fit to data as well because it is the same as a single decision tree. As the number of trees increases, we can see that the generalization error always converges. The equation indicates that the RF may overcome the over-fitting weakness (Trigila et al. 2015; Pourghasemi and Kerle 2016). It also shows excellent performance quality and stability. Therefore, when we add trees to the random forest, the tendency to over-fitting should decrease (thanks to bagging and random feature selection). The statistical package R was used for RF modeling. In landslide susceptibility maps, the model output (between 0 and 1) represents for each grid cell the probability p that is to belong to a landslide.
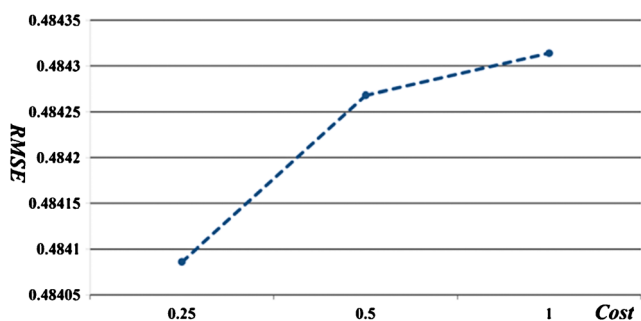

Fig. 9 Parameter training of SVM (radial)

Model evaluation and validation

The prediction models mentioned above is a two-class prediction problem (i.e., binary classification). The results have labeled either as a landslide or as a no-landslide. This study gives a name to landslides as a positive (P) and no-landslides as a negative (N). If the predicted outcomes are P, and the actual value is also P, then it is true positive (TP). However, if the actual value is N, then it is false positive (FP). Conversely, a true negative (TN) occurs both the prediction outcome and the actual value are N, and false-negative (FN) is a situation that the prediction outcome is N, while the actual value is P. The four outcomes can formulate a 2 × 2 confusion matrix (Table 3).

The sensitivity TPR is defined as the proportion of positive cases that have correctly identified according to the equation:

$$TPR = \frac{TP}{P} = \frac{PP}{TP + FN} \quad (11)$$

The specificity FPR is defined as the proportion of negatives cases that have incorrectly classified as positive, according to the equation:

$$FPR = \frac{TP}{N} = \frac{FP}{FP + TN} \quad (12)$$

Researchers consider the receiver operating characteristic (ROC) curve to measure the accuracy of landslide susceptibility mapping (Hong et al. 2016; Pourghasemi and Kerle 2016; Chen et al. 2019). ROC curve is a graph of TPR versus FPR. It depicts the relative trade-offs between the false positive and true positive. Since the TPR is equivalent to sensitivity and FPR is equal to 1-specificity, the ROC graph is sometimes called the sensitivity vs. 1-specificity plot. This curve is referred to as the receiver operating characteristic curve (ROC) for the diagnostic test. The accuracy performance quality of models can be calculated by the area under the ROC (AUC). The higher the AUC value, the higher the accuracy of the model is.

Results

The Jiuzhaigou earthquake landslide is mainly affected by earthquakes with peak acceleration (PGA), and it follows by the distance to the seismic faults (DF), slope degree (SLOPE), normalized difference vegetation index (NDVI), land use types (LT), and relief degree of the land surface (RDL) as shown in Fig. 7.

This study used the VIF to quantify the severity of multicollinearity about the six predisposing factors. This study reveals that when the value of VIF is greater than 5, it indicates that this factor (PGA, DF, SLOPE, NDVI, LT, or RDL) exists

Fig. 10 Parameter training of SVM (polynomial)
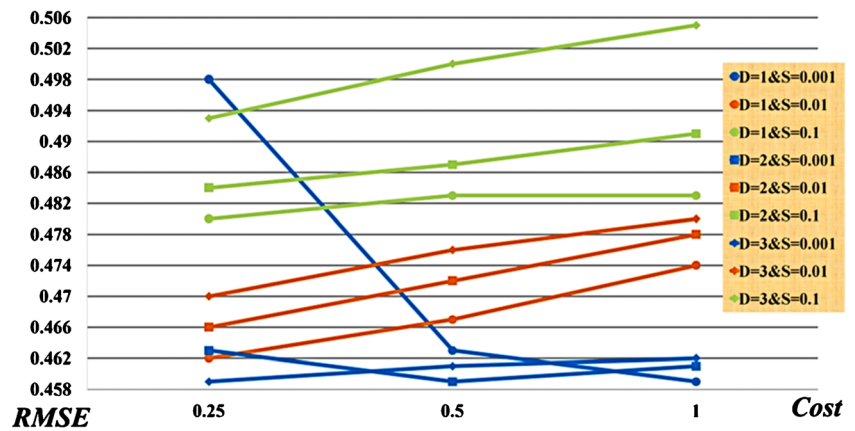
in multicollinearity with other factors. When the value of VIF is greater than 10, it indicates a high degree multicollinearity between those factors in Table 4.

The SVM model (linear and RBF) was mainly controlled by the parameter of the cost (i.e., cost of constraints violation. It is the "C"-constant of the regularization term in the Lagrange formulation). We used 10-fold cross-validation to ten times for training the optimal time consumption. We also used the RMSE to evaluate the accuracy of cost. Training results show that the optimal value of linear is 0.25 and RBF is 0.25 (Fig. 8 and Fig. 9). In the SVM model (polynomial), the parameter of degree (D), scale (S), and cost will control the model. According to the RMSE test, the optimal value of D is 2, S is 0.001, cost is 0.5 (Fig. 10).

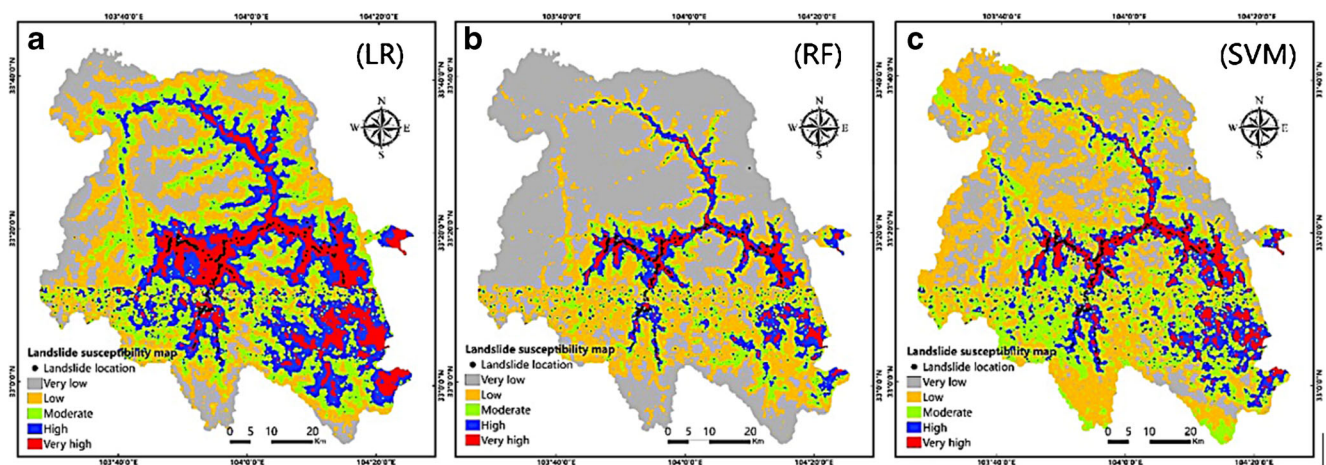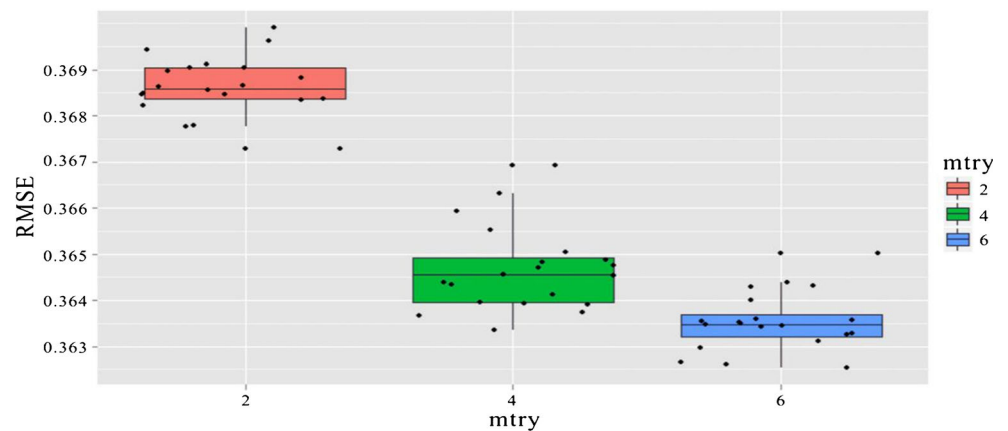**Fig. 11** Boxplot of RMSE values by ten times 10-fold cross-validation

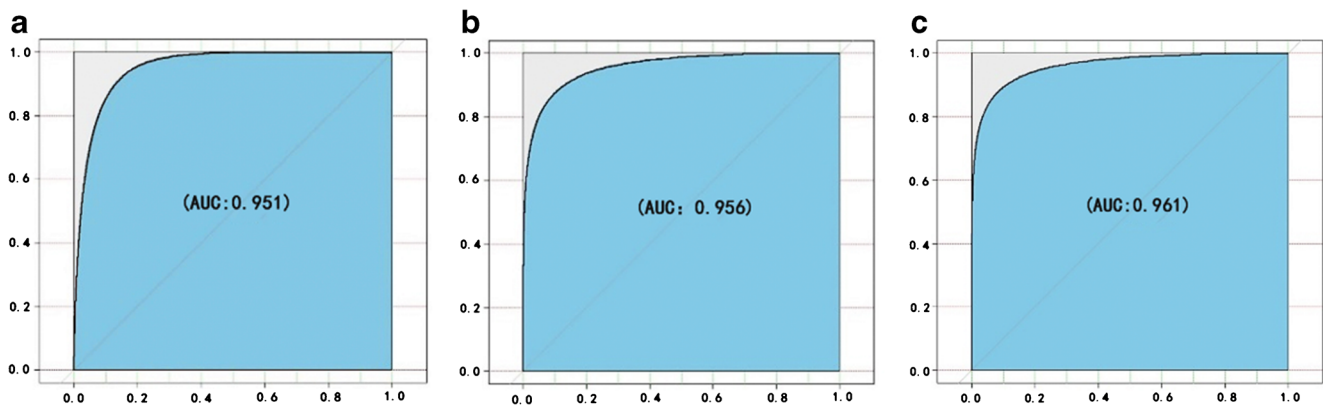**Fig. 12** Landslide susceptibility maps. **a** LR. **b** RF. **c** SVM

Fig. 13 **a** ROC of logistic regression. **b** ROC of SVM (RBF). **c** ROC of random forest

The RF model has a series of decision trees, and it was mainly dominated by the parameter named mtry (a parameter of RF: the number of variables tries to split at each node). In the model construction of RF, we also used the ten times 10-fold cross-validation to train the optimal value of mtry. The training result shows that the optimal value of mtry is 6 (Fig. 11).

Utilizing the three models, we generated three landslide susceptibility maps (Fig. 12).

Discussion

In this study, the ROC of logistic regression is shown in Fig. 13a. The average value of AUC is 0.951. The SVM result's three different kernel functions show that the RBF can perform better than the linear and polynomial. Therefore, in this study, the radial kernel function is most suitable for the landslide susceptibility assessment. The ROC is illustrated in Fig. 13b, and the average value of AUC is 0.956. The random forest model shows that the average value of AUC is 0.961 (Fig. 13c). By comparing with logistic regression (the average is 0.951) and SVM (RBF) (the average is 0.956), we found that the accuracy of random forest is the highest. In other words, the performance quality of RF is better than the other two mentioned models.

And through landslide susceptibility maps, we could know that the areal proportion of the high and very high susceptibility classes is 39.8% for the LR, 11.4% for the RF, and 17.9% for SVM. Obviously, the convergence of RF is higher. The

details of the regional statistics are shown in Table 5 and Table 6.

According to Can et al. (2005), we can consider one factor and then evaluate the accuracy of the landslide susceptibility map. Most of the known landslides appear in the high susceptibility and a very high susceptibility area. The predicted high susceptibility and very high susceptibility areas are as small as possible. Thus, the ratio can be used to evaluate the accuracy:

$$\text{ratio} = \frac{A}{B} \tag{13}$$

where A is the proportion of landslide points that have occurred in the high and very high susceptibility classes. B is the areal proportion of the classes with a high and very high susceptibility.

The LR indicates that 86.9% of the landslides have performed in 48.8% of the high susceptibility zones. The SVM shows that 83.8% of the landslides have appeared in 43.9% of the high susceptibility area. The RF indicates that 79.6% of the landslides have emerged in 38.4% of the high susceptibility area. The ratio of LR is 1.78, SVM with the ratio of 1.9, and RF with a ratio of 2.07 (Fig. 14). It reveals that RF has the best effect in the landslides susceptibility quality assessment.

The model accuracy evaluation and landslide susceptibility mapping show that the RF model is better than the other two models. The RF model can quantify the importance of factors based on the control variable method. The RF turns the single factor into a random number, and the mean decrease accuracy indicates the importance of the factor.

Table 5 The landslide susceptibility area statistics for the three models

Area	LR	RF	SVM
Very low/low	0.1834	0.3640	0.3995
Moderate	0.3276	0.2517	0.1613
High	0.4615	0.3372	0.1776
Very high	0.0274	0.0470	0.2616

Table 6 The fractional distribution of landslide points over the four susceptibility classes

Landslide	LR	RF	SVM
Very low/low	0.0169	0.0477	0.0858
Moderate	0.1144	0.1568	0.0763
High	0.7595	0.5826	0.2182
Very high	0.1091	0.2129	0.6197

Fig. 14 The ratio scores of the three models
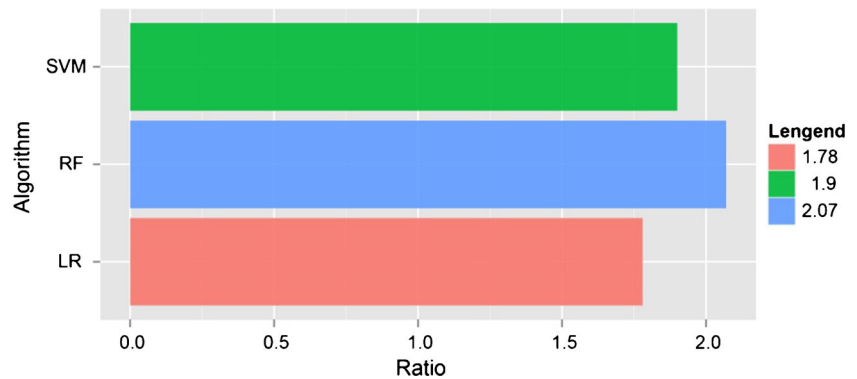
Conclusion

This study selected six impact factors (slope degree, relief degree of the land surface, peak ground acceleration, normalized difference vegetation index, land use types, and distance from faults). We compared three models of landslide susceptibility mapping (i.e., LR, SVM, and RF) for the Jiuzhaigou area. This study has employed the performance of the AUC and the ratio to evaluate the model's accuracy. The result concludes that the average value of AUC is 0.951 in the LR, 0.956 in the SVM, and 0.961 in the RF. It concludes that RF has the best prediction performance. According to the RF model, we can also conclude that the earthquake landslide is affected by the PGA, DF, and SLOPE in the Jiuzhaigou area, which is closely related to the terrain and geological environment.

Many impact factors influence the occurrence of landslides, and each impact factor has a different influence. This study concluded that without considering the lithologic factor, the models still have a high accuracy performance to generate susceptibility maps. It needs further studies to prove that the statistical method could get a great result under the condition of the lack of landslide impact factors. It is far from adequate in aid of preservation efforts for the giant panda habitat, and there is much more to do. Therefore, we suggest taking more geographical locations and areas like Jiuzhaigou in future studies where both landslide-prone areas are and wild giant panda habitats are dominant.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Ahmed B (2015) Landslide susceptibility modelling applying user-defined weighting and data-driven statistical techniques in Cox's Bazar Municipality, Bangladesh. Nat Hazards 79(3): 1707–1737

Ali SA, Pirasteh S (2004) Geological applications of Landsat Enhanced Thematic Mapper (ETM) data and Geographic Information System (GIS): mapping and structural interpretation in south-west Iran, Zagros Structural Belt. Int J Remote Sens 25(21):4715–4727

Ayalew L, Yamagishi H (2005) The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains. Central Japan Geomorphol 65(1–2):15–31

Bourenane H, Bouhadad Y, Guettouche MS (2014) GIS-based landslide susceptibility zonation using bivariate statistical and expert approaches in the city of Constantine (Northeast Algeria). B Eng Geol Environ 74(2):337–355

Brabb EE, Pampeyan EH, Bonilla MG (1972) Landslide susceptibility in San Mateo County, California. Miscellaneous Field Studies Map.

Can T, Nefeslioglu HA, Gokceoglu C (2005) Susceptibility assessments of shallow earthflows triggered by heavy rainfall at three catchments by logistic regression analyses. Geomorphology. 72(1–4):250–271

Cao J, Zhang Z, Wang C, Liu J, Zhang L (2019) Susceptibility assessment of landslides triggered by earthquakes in the Western Sichuan Plateau[J]. Catena 175(175):63–76

Carrara A (1983) Multivariate models for landslide hazard evaluation. Math Geol 15(15):403–426

Chen W, Shahabi H, Shirzadi A, Hong H, Akgun A, Tian Y, Liu J, Zhu AX, Li S (2018) Novel hybrid artificial intelligence approach of bivariate statistical-methods-based kernel logistic regression classifier for landslide susceptibility modeling. BullInt Assoc Eng Geol 78:4397–4419. https://doi.org/10.1007/s10064-018-1401-8

Chen W, Zhao X, Shahabi H, Shirzadi A, Khosravi K, Chai H, Zhang S, Zhang L, Ma J, Chen Y, Wang X, Bin Ahmad B, Li R (2019) Spatial prediction of landslide susceptibility by combining evidential belief function, logistic regression and logistic model tree. Geocarto Int 34: 11–1201. https://doi.org/10.1080/10106049.2019.1588393

Dai FC, Lee CF, Ngai YY (2002) Landslide risk assessment and management: an overview. Eng Geol 64(1):65–87

Dou J, Yamagishi H, Pourghasemi HR (2015) An integrated artificial neural network model for the landslide susceptibility assessment of Osado Island, Japan. Nat Hazards 78(3):1–28

Elayaraja S, Ganapathy GP (2015) Evaluation of seismic hazard and potential of earthquake-induced landslides of the Nilgiris, India. Nat Hazards 78(3):1997–2005. https://doi.org/10.1007/s11069-015-1816-5

Erener A, Düzgün HSB (2012) Landslide susceptibility assessment: what are the effects of mapping unit and mapping method? Environ Earth Sci 66:859–877

Faris F, Wang F (2014) Investigation of the initiation mechanism of an earthquake-induced landslide during rainfall: a case study of the Tandikat landslide, West Sumatra, Indonesia. Geoenviron Disast 1(1):1–18

Farrokhnia A, Pirasteh S, Pradhan B, Pourkermani M, Arian M (2011) A recent scenario of mass wasting and its impact on the transportation in Alborz Mountains, Iran: contribution from Geo information technology. Arab J Geosci 4:1337–1349

Feng Z, Tang Y, Yang Y, Zhang D (2008) Relief degree of land surface and its influence on population distribution in China. J Geogr Sci 18(2):237–246

Hong H, Pourghasemi HR, Pourtaghi ZS (2016) Landslide susceptibility assessment in Lianhua County (China): a comparison between a random forest data mining technique and bivariate and multivariate statistical models. Geomorphology. 259:105–118

Hosseinalizadeh M, Karaminejad N, Rahmati O, Keesstra S, Alinejad M, Behbahani AM (2019) How can statistical and artificial intelligence approaches predict piping erosion susceptibility? Sci Total Environ 646:1554–1566

Jaafari A, Najafi A, Pourghasemi HR et al (2014) GIS-based frequency ratio and index of entropy models for landslide susceptibility assessment in the Caspian forest, northern Iran. Int J Environ Sci Technol I 11(4):909–926

Juliev M, Mergili M, Mondal I, Nurtaev B, Pulatov A, Hübl J (2019) Comparative analysis of statistical methods for landslide susceptibility mapping in the Bostanlik District, Uzbekistan. Sci Total Environ 653:801–814

Kaya A, Alemdağ S, Dağ S (2015) Stability assessment of high-steep cut slope debris on a landslide (Gumushane, NE Turkey). Bull Eng Geol Environ 75(1):89–99

Lai J-S, Tsai F (2019) Improving GIS-based landslide susceptibility assessments with multi-temporal remote sensing and machine learning. Sensors 19(17):3717. https://doi.org/10.3390/s19173717

Lee S, Talib JA (2005) Probabilistic landslide susceptibility and factor effect analysis. Eng Geol 47(7):982–990

Lee S, Ryu JH, Min K, Won JS (2003) Landslide susceptibility analysis using GIS and artificial neural network. Earth Surf Process Landf 28(12):1361–1376

Lee CT, Huang CC, Lee JF (2008) Statistical approach to earthquake-induced landslide susceptibility. Eng Geol 100(1–2):43–58

Li YS, Huang C, Yi SJ et al (2017a) Study on seismic fault and source rupture tectonic dynamic mechanism of Jiuzhaigou Ms7.0 earthquake. J Eng Geol 25(4):1141–1150 (In Chinese)

Li L, Liu R, Pirasteh S, Chen X, He L, Li J (2017b) A novel genetic algorithm for optimization of conditioning factors in shallow translational landslides and susceptibility mapping. Arab J Geosci 10(9). https://doi.org/10.1007/s12517-017-3002-4

Li X, Ling S, Sun C et al (2019) Integrated rockfall hazard and risk assessment along highways: an example for Jiuzhaigou area after the 2017 Ms 7.0 Jiuzhaigou earthquake, China. J Mt Sci 16(6):1318–1335

Liu HL, Ma YB, Zhang WG et al (2020) Summary of the application of big data technology in geological disaster prevention. J Disast Prevent Mitigat Eng:1–13

Meten M, Bhandary NP, Yatabe R (2015) GIS-based frequency ratio and logistic regression modelling for landslide susceptibility mapping of Debre Sina area in Central Ethiopia. J Mt Sci 12(6):1355–1372

Pandit SN, Maitland BM, Pandit LK, Poesch MS, Enders EC (2017) Climate change risks, extinction debt, and conservation implications for a threatened freshwater fish: carmine shiner (Notropis percobromus). Sci Total Environ 598:1–11

Pawluszek K, Borkowski A, Tarolli P (2018) Sensitivity analysis of automatic landslide mapping: numerical experiments towards the best solution. Landslides 15:1851–1865

Pham BT, Tien Bui D, Prakash I, Dholakia MB (2017) Hybrid integration of multilayer perceptron neural networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. CATENA 149:52–63

Pirasteh S, Shamsipur G, Liu G, Zhu Q, Ye C (2020) Developing landslide deformation geometric algorithm for modelling and simulation incorporating LiDAR-derived DEMs and UAV. Earth Science Informatics. pp: 1-15, https://doi.org/10.1007/s12145-019-00437-5.

Pourghasemi HR, Kerle N (2016) Random forests and evidential belief function-based landslide susceptibility assessment in Western Mazandaran Province, Iran. Environ Earth Sci 75:1–17

Reichenbach P, Rossi M, Bruce DM, Mihir M, Guzzettia 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

Shirani K, Pasandi M, Arabameri A (2018) Landslide susceptibility assessment by Dempster–Shafer and Index of Entropy models, Sarkhoun basin, Southwestern Iran. Nat Hazards 6:1–40

Torgoev A, Havenith HB (2013) Parametric numerical study of seismic slope stability and verification of the Newmark method. In: Earthquake-Induced Landslides. Springer, Berlin Heidelberg. https://doi.org/10.1007/978-3-642-32238-9_68

Trigila A, Iadanza C, Esposito C, Scarascia-Mugnozza G (2015) Comparison of logistic regression and random forests techniques for shallow landslide susceptibility assessment in Giampilieri (NE Sicily, Italy). Geomorphology 249:119–136

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

Wan S (2009) A spatial decision support system for extracting the core factors and thresholds for landslide susceptibility map. Eng Geol 108(3-4):237–251

Wang KL, Lin ML (2010) Development of shallow seismic landslide potential map based on Newmark's displacement: the case study of Chi-Chi earthquake. Taiwan Environ Earth Sci 60(4):775–785

Wang J, Jin W, Cui YF, Zhang WF, Wu CH, Alessandro P (2018) Earthquake-triggered affecting a UNESCO Natural Site: the 2017 Jiuzhaigou Earthquake in the World National Park, China. J Mt Sci 15(7):1412–1428

Wang Y, Wu X, Chen Z, Ren F, Feng L, Du Q (2019) Optimizing the predictive ability of machine learning methods for landslide susceptibility mapping using SMOTE for Lishui City in Zhejiang Province, China. Int J Environ Res Public Health 2019(16):368. https://doi.org/10.3390/ijerph16030368

Wei-Le L, Huang RQ, Qiang X (2013) Rapid prediction of co-seismic landslides triggered by Lushan earthquake, Sichuan, China. J Chengdu Univ Technol 40(3):264–274. https://doi.org/10.3969/j.issn.1671-9727.2013.03.06

Wu S, Jin Y, Zhang Y (2004) Investigations and assessment of the landslide hazards of Fengdu county in the reservoir region of the Three Gorges project on the Yangtze River. Environ Geol 45(4):560–566

Xu C, Xu X, Yu G (2013) Landslides triggered by slipping-fault-generated earthquake on a plateau: an example of the 14 april 2010, ms 7.1, yushu, china earthquake. Landslides 10(4):421–431

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. 327:170–187

Yang L, Nollen N, Ahluwahlia JS (2015) An application in identifying high-risk populations in alternative tobacco product use utilizing

logistic regression and CART: a heuristic comparison. BMC Public Health 15(1):1–9

Ye C, Li Y, Liang L, Pirasteh S, Cui P, Li J (2019) Landslide detection of hyperspectral remote sensing data based on deep learning with constrains. IEEE J Select Topic Appl Earth Observ Remote Sens 12: 5047–5060. https://doi.org/10.1109/JSTARS.2019.2951725

Zhang Y, Zhang G, Hetland EA et al (2018) Source fault and slip distribution of the 2017 Mw 6.5 Jiuzhaigou, China, earthquake and its tectonic implications. Seismol Res Lett 89(4):1345–1353

Zhou S, Fang L (2015) Support vector machine modeling of earthquake-induced landslides susceptibility in central part of Sichuan province, China. Geoenviron Disast 2(1):1–12

Zhu Q, Chen L, Hu H, Pirasteh S, Li H, Xie X (2020) Unsupervised feature learning to improve transferability of landslide susceptibility representations. IEEE J Sel Top Appl Earth Obs Remote Sens 13: 2917–3930. https://doi.org/10.1109/JSTARS.2020.3006192