




# Landslide hazard assessment based on Bayesian optimization–support vector machine in Nanping City, China

Wei Xie<sup>1,2</sup> · Wen Nie<sup>1,2,3</sup>  · Pooya Saffari<sup>2</sup> · Luis F. Robledo<sup>4</sup> · Pierre-Yves Descote<sup>4</sup> · Wenbin Jian<sup>5</sup>

Received: 18 January 2021 / Accepted: 8 June 2021 / Published online: 15 June 2021  
© The Author(s), under exclusive licence to Springer Nature B.V. 2021

## Abstract

Landslide hazard assessment is critical for preventing and mitigating landslide disasters. The tuning of hyperparameters is of great importance to achieve better accuracy in a landslide hazard assessment model. In this study, a novel approach is proposed for landslide hazard assessment with support vector machine (SVM) as the primary model and Bayesian optimization (BO) algorithm as the parameter tuning method. This study describes 1711 historical landslide disaster points in Nanping City, and a total of 12 landslide conditioning factors including elevation, slope, aspect, curvature, lithology, soil type, soil erosion, rain-fall, river, land use, highway, and railway were selected. The multicollinearity diagnosis was performed on the factors using the Spearman correlation coefficient. For model validation, 1711 landslides and 1711 non-landslides were collected as the dataset and divided into a training dataset (50%) and a testing dataset (50%). The performance of the model was evaluated by the confusion matrix and receiver operating characteristic (ROC) curve. The results of the confusion matrix accuracy and the area under the ROC curve showed that the BO-SVM model (89.53%, 0.97) performed better than the SVM model (84.91%, 0.93). In addition, the landslide hazard maps generated by the BO-SVM model had better overall results than that by the SVM model.

**Keywords** Landslide hazard assessment · Bayesian optimization method · Support vector machine · GIS · Machine learning

---

✉ Wen Nie  
wen.nie@vip.tom.com

<sup>1</sup> School of Resources and Environmental Engineering, Jiangxi University of Science and Technology, 341000 Ganzhou, People's Republic of China

<sup>2</sup> Spatial Information Technology and Big Data Mining Research Center in Southwest Petroleum University, 610500 Chengdu, People's Republic of China

<sup>3</sup> Zhejiang Zhipu Engineering Technology Co., Ltd, 313002 Huzhou, People's Republic of China

<sup>4</sup> Universidad Andres Bello, Santiago 7500971, Chile

<sup>5</sup> Department of Geotechnical and Geological Engineering, Fuzhou University, 350108 Fuzhou, People's Republic of China

## 1 Introduction

Landslides are geological hazards that slide down rock or soil along a certain weak surface under the action of gravity (Fan et al. 2019; Hungr et al. 2013; Qiu et al. 2019). Every year, landslide disasters are responsible for a large number of casualties and economic losses worldwide (Kirschbaum et al. 2009; Froude et al. 2018). The aggravated change in climate has unequivocally affected the stability of natural and engineered slopes, posing greater risk of landslides (Gariano et al. 2016). Therefore, mitigating the serious threat of landslide disasters and preventing new landslide disasters have become an increasingly important issue to address (Intrieri et al. 2019).

Landslide hazard assessment is an effective measure to prevent landslide hazards. This method can provide key information for disaster prevention, disaster mitigation, and disaster risk reduction (Westen et al. 2008; Xu et al. 2012).

In recent years, with the rapid development of geographic information systems and artificial intelligence (AI) technologies, a large number of landslide hazard assessment methods use a variety of advanced algorithms and models (logistic regression, support vector machines, Bayesian methods, decision tree methods, artificial neural networks, among others) (Jafarian et al. 2019; Olen et al. 2018; Theron et al. 2018; Violante et al. 2018; Wu et al. 2018; ). Bourenane et al. (2016) used the frequency ratio and logistic regression to develop a landslide hazard map in Constantine city. Xie et al. (2021) developed a machine learning cluster containing multiple machine learning methods for landslide susceptibility mapping, and the results showed that the method is effective in assessing such landslide susceptibility. Besides, an increasing number of literature reports indicate that integrated machine learning methods can avoid the shortcomings of a single approach Chen et al. 2021a; Pal et al. 2019).

In general, GIS provides a platform for collecting, organizing, and analyzing landslide events and landslide conditioning factors. Machine learning (ML) techniques provide additional solutions for calculating the relationship between the landslide conditioning factors and landslide events (An et al. 2018; Chen et al. 2021; Moresi et al. 2020; Pal et al. 2019). Among the ML techniques, support vector machine (SVM) is one of the most used methods, and it has performed satisfactorily in several previous studies. Marjanovic et al. (2011) compared SVM, decision trees, and logistic regression in a specific area of the Fruska Gora Mountain (Serbia), indicating that the SVM classifier outperformed the other methods. Chen et al. (2017b) used the maximum entropy, SVM, and artificial neural network (ANN) to find the “Spatial contraindication” pattern by their ensembles. SVM is the most practical model with the highest spatial area in highly susceptible classes. Luo et al. (2019) applied ANN, SVM, and information value model to assess a mining landslide sensitivity analysis. The ANN model and SVM achieved high prediction capability, proving their advantage of solving nonlinear and complex problems. In addition, Wang et al. (2019) reported good results obtained by the SVM method using support vector regression for short-term traffic flow prediction in the problems of classification and regression.

However, the hyperparameter selection of SVM is often confusing and affects the precision and generalization ability of the model (Dou et al. 2020; Zhao et al. 2020). For SVM, the kernel function type is the most important hyperparameter, and the penalty factor (C) and Gamma also affect the performance of the model. Many studies used default hyperparameters in software or tools, resulting in less than optimal outcomes (Abdollahi et al. 2018; Chen et al. 2017b). Some studies used random search, grid search, or genetic algorithms to optimize the hyperparameters (Luo et al. 2019;

Tang et al. 2019). Nevertheless, these optimization methods have clear drawbacks that affect landslide hazard assessment. Both random search and grid search are blind, thus consuming a lot of time. Genetic algorithms tend to fall into local optimality, and the overall performance is compromised.

The hyperparameter selection of SVM was optimized by the Bayesian optimization (BO) method, and a new BO-SVM model also developed in this study. To evaluate the optimization effect of the BO-SVM method, this study compared the performance of this method with the common SVM and random search for landslides assessment in Nanping City of China.

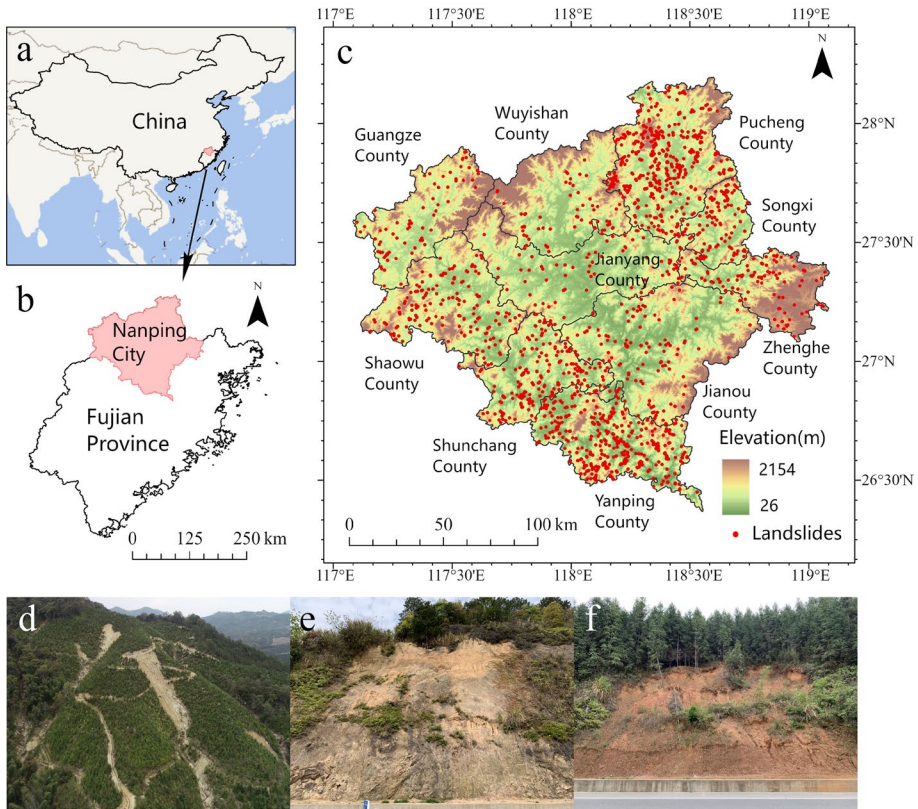
## 2 Study area and data

### 2.1 Study area and landslide inventory

The study area is Nanping city, located in the northwestern part of Fujian province, China. The geographic coverage of the study area is 117°46'31"–118°17'9"N latitude and 42°14'55"–42°48'30"E longitude, with an area of ~26,300 km<sup>2</sup> (Fig. 1). Nanping is administratively divided into 10 counties and districts including Jianyang, Yanping, Shunchang, Pucheng, Guangze, Songxi, Zhenghe, Shaowu, Wuyishan, and Jian'ou (Fig. 1).

The elevation ranges from 20 to 2150 m above the sea level, and the land above 1000 m accounts for 12% of the total area. The low hills are widely distributed in this whole area, and high mountains are only in the northeast and southwest. Mountain basin valleys are distributed alternately along the river. Fault-block mountains are structurally dominated by faults with steep peaks. Surface topographical features in this study area are strongly affected by tectonic movement. This study area is geotectonically located along the southeastern edge of the Eurasian continental plate and bordering the Pacific plate, which is the most active area of Cenozoic tectonic-magmatic activity in the Pacific Rim. The geological structure of this study area is complex with obvious tectonic and geomorphic features. The climate type of the study area is subtropical monsoon climate. The average annual rainfall ranges from approximately 1500 mm to 2200 mm, and it is concentrated in summer and tropical cyclones, usually causing considerable precipitation. The main rivers in the research area are the Minjiang River, the Jian River, and the Futun River.

Nanping is one of the most landslide-prone areas of southeastern China with numerous historical landslide events (Zhang et al. 2018). The typhoon rainstorms and the interaction of anthropic activities and engineering geological conditions are the main factors causing landslides in this study area (Yin et al. 2013). The landslide inventory in this study area was collected by a field geological survey, as shown in Fig. 1. In this study, the term "landslide" includes slide, fall, and flow. The classification comes from the updated Varnes landslide classification (Hung et al. 2013), and all landslides are represented by spatial points. There were 1008 cases of slides, 679 cases of falls, and 24 cases of flows. In terms of trigger, most of the landslides were caused by rainfall (about 80%); Fig. 1d shows the landslide with obvious signs of rainfall. The remaining landslides are caused by human engineering activities and groundwater. 65% of these landslides are shallow soil landslides, and the rest are rock collapses (Fig. 1e) with a small number of deep landslides. The surface of this study area is covered by a large amount of red and sandy soil, being the engineering properties of these soils very poor (Fig. 1f).



**Fig. 1** Location of the study area and landslide inventory. **a** The location of the study area in China; **b** The location of the study area in Fujian Province; **c** landslide inventory. The red dot refers to 1711 historical landslides studied in our case; **d–f** photographs of different types of landslide cases, with a landslide induced by rainfall **d**, a rock collapses **e** and a shallow landslide **f**

## 2.2 Landslide conditioning factors

According to the geographic environment configuration and data availability of the study area, 12 landslide conditioning factors: elevation, slope, aspect, curvature, lithology, soil type, soil erosion, precipitation, river, land cover, highway, and railway (Fig. 2) were selected. The 12 input thematic variables were classified into five clusters: (I) morphological (4 variables), (II) geological (3 variables), (III) hydrological (2 variables), (IV) land cover (1 variable), and (V) anthropological activities (2 variables) (Table 1). All the factors were discretized into categorical variables. For category factors, their original structure such as lithology, land use, soil type, and soil erosion was retained. The aspect factor was classified according to the direction, as shown in Fig. 2b. The remaining continuous factors were classified according to the natural break method.

Topography and landform play an important role in controlling the formation of landslides (Ambrosi et al. 2018). Digital elevation model (DEM) and slope degree were widely used factors and particularly effective in landslide predicting (Reichenbach et al. 2018). The elevation (Fig. 1c) has a direct effect on human engineering activities and other

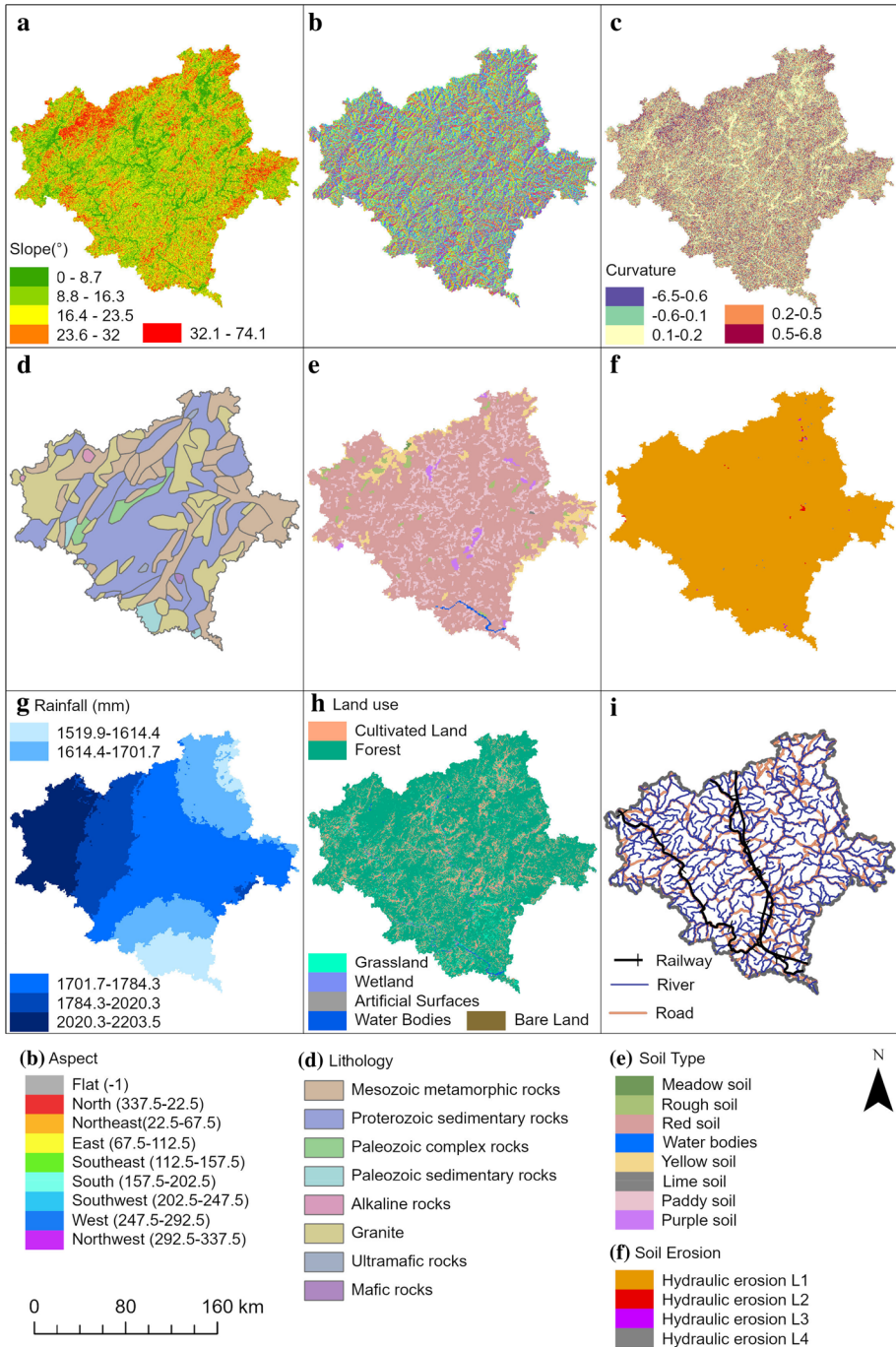
**Table 1** Landslide conditioning factors

Cluster	Factor	Description	Source
Morphological	Elevation	ASTER GDEM V2, 30 m resolution	<a href="http://www.gscloud.cn/">http://www.gscloud.cn/</a>
	Slope	30 m resolution	Extracted by DEM
	Aspect	30 m resolution	Extracted by DEM
	Curvature	30 m resolution	Extracted by DEM
Geological	Lithology	Vector data	<a href="http://www.geodata.cn">http://www.geodata.cn</a>
	Soil type	Reclassify to 30 m resolution	<a href="http://www.resdc.cn">http://www.resdc.cn</a>
	Soil erosion	Reclassify to 30 m resolution	<a href="http://www.resdc.cn">http://www.resdc.cn</a>
Hydrological	Rainfall	Average rainfall from 1980 to 2015, interpolated from rainfall site data	<a href="http://data.cma.cn/">http://data.cma.cn/</a>
	River	Vector data	<a href="http://www.geodata.cn">http://www.geodata.cn</a>
Land cover	Land use	30 m resolution	<a href="http://www.webmap.cn">http://www.webmap.cn</a>
Anthropological activity	Highway	Vector data	<a href="http://www.webmap.cn">http://www.webmap.cn</a>
	Railway	Vector data	<a href="http://www.webmap.cn">http://www.webmap.cn</a>

environmental factors, thereby affecting the stability of a slope. The gravitational potential energy of the slope is greater at higher elevations relative to lower ones. The elevation also determines the land cover, anthropogenic activities, and climate type (Chen et al. 2020). The slope degree (Fig. 2a) affects the stability and overall movement rate of the unstable rock and soil on the slope (Lo et al. 2018) and is mainly in the range 6–20° in the selected study area. Usually, the greater the slope degree, the worse the stability of the slope. The aspect determines the direction of solar radiation and water flow (Fig. 2b); the curvature (Fig. 2c) affects the acceleration and deceleration of flow, convergence, and dispersion (Youssef et al. 2015).

Geological structure and soil properties (Fig. 2d–f) would directly predict the occurrence of landslide and its mechanism (Zeze et al. 2017). In stratigraphic lithology, landslides often occur in soft structural planes and weak rock layers. Proterozoic sedimentary rocks occupy about half of the selected study area, followed by Granite and Mesozoic metamorphic rocks (Fig. 2d). Mesozoic metamorphic rocks, Paleozoic complex rocks, Paleozoic sedimentary rocks, alkaline rocks, and mafic rocks are identified within the study area. The sedimentary rocks exposed in the study area are laminated and contain a large amount of debris, whereas those in some of the study areas contain a large amount of clay, with poor engineering properties, and a large number of landslides occur in these areas. For soil properties, looser soil and a greater degree of soil erosion are conducive to landslide breeding (Sorbino et al. 2009). Red and paddy soil are the main soil types in the study area. The degree of hydraulic erosion increases from L1 to L4, and Fig. 2f shows that most of the study area suffers from mild hydraulic erosion. The soil erosion classification referred to the People’s Republic of China industry standard SL190-96 “*Soil erosion classification and classification standards*”.

The hydrological conditioning factors and changes in the land cover (Fig. 2g, h) are predisposing factors or direct factors of landslides (Phong et al. 2019). Dozens of landslides occur in the rainy season every year in this study area, indicating that hydrology has a great effect on landslides. Continuous rainfall can directly cause landslides and erode the slope leading to instability. For the land cover, bare slopes are more prone to instability than slopes with lush vegetation, and forests with luxuriant root systems are more stable than



**Fig. 2** Landslide conditioning factors: **a** slope; **b** aspect; **c** curvature; **d** lithology; **e** soil type; **f** soil erosion; **g** rainfall; **h** land use; **i** railway, river, and highway

grasslands. The forest coverage rate in the study area is 74.75%, and the remaining lands are mainly cultivated lands and grasslands.

The direct effects of anthropological activities on slopes are increasing with the development of economic engineering. A large number of infrastructures and road constructions have destroyed the original structure of hillsides, aggravating the slope instability. The construction of railways (Fig. 2i) and highways (Fig. 2i) in the mountainous areas usually involves the excavation of tunnels and manual cutting of slopes. Therefore, the areas along highways and railways are the worst-hit areas of landslides and the key areas for disaster reduction and prevention (Eeckhaut et al. 2010).

### 3 Methods

#### 3.1 Mapping units and dataset division

In most current studies, the mapping units commonly used in landslide susceptibility mapping and landslide hazard assessment are grid units, geomorphic units, administrative units, unique condition units, and slope units (Reichenbach et al. 2018). Compared with other units, grid units usually perform better for complex calculations and simulation processes (Yang et al. 2019a). Therefore, in this study, the grid units were selected as the basic mapping units. In consideration of the total study area and the computational complexity, a 300 m × 300 m grid was selected, resulting in a total number of 374,666 units. The number of landslide points falling into each unit was calculated separately. The unit with several 0 was recorded as 0, while the unit with a number rather than 0 was recorded as 1 to form a binary distribution. They are the dependent variables of the hazard assessment model. Then, the attribute values of the landslide factors of each grid cell were spatially overlapped as the independent variables of the model.

In this study, 1711 landslide disaster points fell into 1,653 units and were used as positive samples in the data set. An equal number of non-landslides were chosen as negative samples in an area 300 m away from the landslide points. Therefore, a total of 3306 samples were used for model training (50%) and testing (50%).

#### 3.2 Multicollinearity diagnosis

Feature selection is a necessary step in the process of machine learning modeling and is used to eliminate redundant factors and retain useful factors. Multicollinearity is usually used as an indicator of feature selection, indicating that there might be correlations between multiple conditioning factors (Lee et al. 2018). The existence of multicollinearity would make it difficult to capture useful information from the model, thereby affecting the evaluation results (Yanar et al. 2020). Multicollinearity diagnosis of factors and elimination of redundant factors have a positive effect on the evaluation model. In this study, Spearman correlation analysis was used to analyze each factor in the study area, and the multicollinearity of factors was measured through the correlation coefficient,  $R$ . The  $R$  value range is  $[-1, 1]$  (when  $R > 0$ , the factors are positively correlated,  $|R|$  closer to 1, the higher the correlation; when  $R < 0$ , the factors are negatively correlated; when  $R = 0$ , there is no linear correlation).

### 3.3 Landslide hazard assessment model

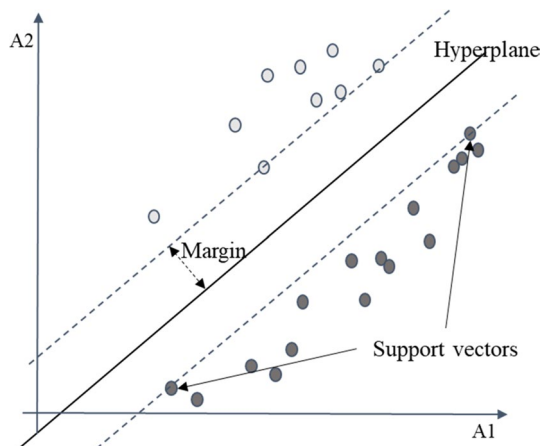
SVM is a machine learning method based on statistical theory, and it integrates multiple techniques such as relaxation variables, maximum interval hyperplane, and kernel function. It is suitable to solve the classification problems of small samples, nonlinearity, and high dimensionality (Cortes et al. 1995). With the development of multidisciplinary integration, SVM was gradually applied to the field of natural disasters. The basic principle is to map the samples of the input space to a high-dimensional characteristic space through nonlinear transformation, followed by determining the optimal classification plane that linearly separates the samples in the characteristic space (Chang et al. 2011; Smola et al. 2004). In the studies of landslide susceptibility assessment and risk assessment, the occurrence of landslides fits well with the characteristics of the algorithm of solving binary classification problems (Ballabio et al. 2012; Tien Bui et al. 2018; Xie et al. 2021).

The schematic diagram of SVM's principle is shown in Fig. 3. The distance between the hyperplane and the nearest sample point is called the margin. The larger the margin, the higher the generalization ability of the classifier. Therefore, the purpose of SVM is to find the hyperplane that maximizes the margin, i.e., the optimal hyperplane. All the points on the hyperplane on both sides of the margin are called support vectors, and the classification boundary is determined only by the support vectors not by other data and the amount of data. Therefore, the adjustment of hyperparameters is extremely critical to the performance of SVM. The main hyperparameters involved in SVM are kernel type,  $C$ , and  $\gamma$ . As mentioned above, the kernel maps the observations into some feature space. Hyperparameter  $C$  controls the trade-off between the decision boundary and accuracy by adding a penalty for each misclassified data point.  $\gamma$  is a parameter related to  $C$  in some kernel types. If  $\gamma$  is large, the effect of  $C$  becomes negligible. If  $\gamma$  is small,  $C$  affects the model in a similar way as it affects a linear model. In this study, the *scikit-learn* package based on Python for SVM implementation was used (Pedregosa et al. 2011).

### 3.4 Bayesian optimization algorithm

The process of implementing machine learning algorithms usually needs to consider the tuning of learning parameters and model hyperparameters (Snoek et al. 2012). The

**Fig. 3** Principle of support vector machine (SVM)





hyperparameters define the attributes of the model or the training process, which have a significant effect on the final effect of the model (Greenhill et al. 2020). BO is a hyperparameter optimization (selection) method from general machine learning algorithms. BO algorithm is widely utilized in the field of cutting-edge artificial intelligence with obvious advantages over genetic algorithm, particle swarm optimization algorithm, or other algorithms (Greenhill et al. 2020; Koblíha et al. 2006). It is a parameter optimization method based on Gaussian process and Bayesian theorem and builds a surrogate for the objective and quantifies the uncertainty in that surrogate using a Bayesian machine learning technique and Gaussian process regression, and followed by using an acquisition function defined from this surrogate to decide the sample location (Frazier 2018). Generally, the problem scenarios that BO algorithm mainly faces are:

$$X^* = \arg_{x \in S} \max f(x) \quad (1)$$

where  $S$  is the candidate set of  $x$ . The goal is to choose  $x$  from  $S$  such that the value of  $f(x)$  is the smallest or largest.

As a sequence optimization problem, BO needs to select an optimal observation value at each iteration. This key problem is perfectly solved by the abovementioned Gaussian process. as expressed by the following formula:

$$f(x) \sim GP(\mu(x), k(x, x^*)) \quad (2)$$

where  $\mu(x)$  is the mean function, and  $k(x, x^*)$  is the kernel function. The form of the Gaussian kernel function is as follows:

$$k(x, x^*) = \exp\left(-\frac{1}{2}\|x - x^*\|^2\right) \quad (3)$$

The hyperparameter value obtained by the BO algorithm replaces the original value. Then, a new hybrid model (BO-SVM) was constructed. Package *hyperopt* on the Python platform was used to implement BO algorithm in this study.

### 3.5 Model evaluation and verification

#### 3.5.1 Confusion matrix

The confusion matrix measures the accuracy of a classifier classification and is also known as the error matrix. It is often used to evaluate the results of binary regression models such as logistic regression and SVM and method can quantitatively express the correct rate of 0-value prediction, the correct rate of 1-value prediction, and the overall prediction rate in the model results (Yang et al. 2019a).

#### 3.5.2 ROC curve

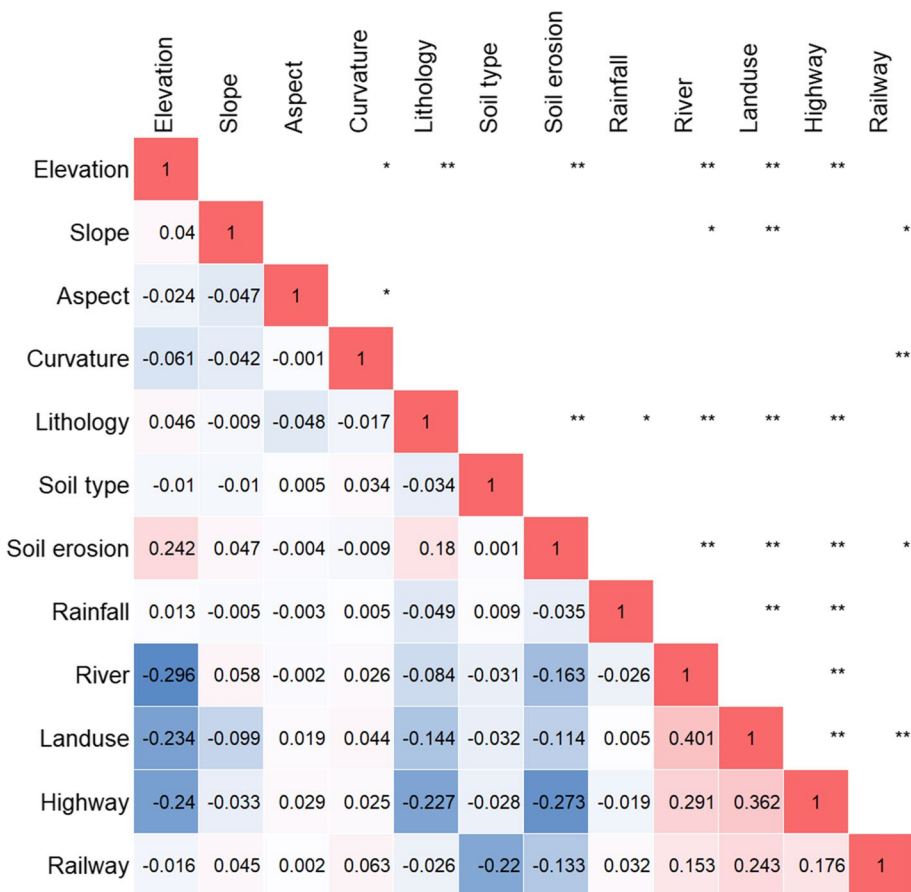
The receiver operating characteristic (ROC) curve is a comprehensive indicator of response sensitivity and specific variables (Chen et al. 2021b; Tehrani et al. 2015). In the landslide risk assessment, the X-axis of the ROC curve specificity indicates the probability of misprediction of the non-disaster points. The Y-axis is the sensitivity, representing the prediction success rate of the disaster point. The prediction accuracy of the model is expressed by the size of the area enclosed by the curve and the abscissa (Chen et al. 2021c). The closer

the curve is to the upper left corner, the higher the accuracy of the model. The area under the curve is called AUC, and the range of AUC values is [0, 1]. The value of AUC closer to 1 indicates the higher accuracy of the model.

## 4 Results

### 4.1 Multicollinearity of factors

Generally, factors with high multicollinearity values should be removed or detected iteratively to ensure the reliability of the model. The multicollinearity diagnosis results among the 12 conditioning factors are presented in Fig. 4. The correlation coefficient between each implemented factor is less than 0.5, indicating low multicollinearity between factors. Consequently, in this study, all 12 conditioning factors were retained. The value of collinearity among most factors is around 0, indicating extremely low correlation between them.



**Fig. 4** Correlation coefficient of 12 conditioning factors. \* indicates at the significant level  $\alpha=5\%$ , the correlation is statistically significant, \*\*means at the significant level  $\alpha=1\%$ , the correlation is statistically significant

Moreover, the multicollinearity between land use factors and human activities factors has a higher value.

## 4.2 Verification and comparison of models

The hyperparameters corresponding to the most optimal evaluation value of BO proceeding were the radial basis function (RBF), which was used as the kernel; the penalty factor C, which was  $1 \cdot 10^{8.475}$ , and the RBF gamma value, which was  $2.895 \cdot 10^{-7}$ . These values would be set as the hyperparameter values for BO-SVM before modeling, while the SVM used the default hyperparameters. The performance of the SVM and BO-SVM models was verified and compared using the confusion matrix and ROC curve, respectively. ROC curve presses the predictive capabilities of the models, and the confusion matrix represents the details of the predictive ability of the model.

The results of the confusion matrix are as shown in Table 2, indicating the accuracy of BO-SVM as 89.63 %, which is approximately 5 % higher than the SVM with 84.91 %. Compared with SVM, BO-SVM has higher prediction accuracy for landslides events. The prediction accuracy of SVM for landslides events are 88.64 and 81.18 %, respectively, indicating that the prediction accuracy of BO-SVM for negative and positive is relatively robust.

The ROC curve and the area under the ROC curve (AUC) are illustrated in Fig. 5. Figure 5a shows the distribution of the AUC values of the two models with progression of the iterative process. Figure 5b shows the ROC curves and AUC values of the two models on the testing dataset. Generally, the AUC values greater than 0.9 are considered excellent (Merghadi et al. 2020). In this case, both models have high AUC values of more than 0.9, and BO-SVM with 0.97 is 4 % higher than SVM with 0.93. From the iterative process, the BO-SVM has a better distribution of AUC values and most of them are concentrated at the top of the graph. In contrast, the values of SVM are evenly dispersed in the middle and at the top. The iterative trend shows that the BO-SVM is in a state of continuous increment, while SVM has almost no trend of change. The confusion matrix and ROC curve results indicate better performance of the BO-SVM performance than that of the SVM.

**Table 2** The confusion matrix of the SVM and BO-SVM

Method	Landslide occurred(Actual)	Prediction		Percent (%)	Accuracy (%)
		Yes	No		
SVM	Yes	1466	187	88.64	84.91
	No	311	1342	81.18	
BO-SVM	Yes	1523	130	92.14	89.53
	No	216	1437	86.93	

## 4.3 Landslide hazard map

According to the results of the landslide hazard model, the landslide hazard index in the study area was obtained by carrying out the spatial overlay analysis of each conditioning factor. The results ranged from 0 to 1 and were divided into four zones with an interval of 0.25, namely low, moderate, high, and very high (Fig. 6). The two maps show a similar

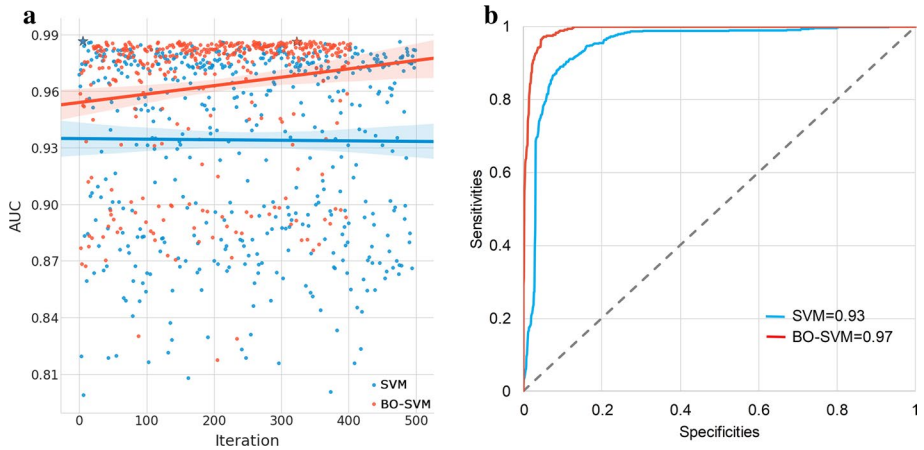
spatial distribution to some extent: the very high hazardous areas are clustered at the top and bottom of the map, and the other areas are relatively low risk hazard. The statistic results of the map units and landslides in each hazard zone are as listed in Table 3. In the BO-SVM results, the low hazard zone area accounted for 54.98 % of the study area, and the number of historical landslides only accounted for 5.96 % of the total landslides. The very high and high hazard zone area accounted for 14.83 % of the total area, and the number of historical landslides accounted for 64.87 % of the total landslides. In contrast, the SVM shows less accurate results, where the low zone and high zone risk areas are similar to that obtained by BO-SVM, while the very high zone risk area only includes 15.14 % landslides. Besides, there are 39.45 % landslides in the moderate zone, demonstrating low confidence for the SVM model.

## 5 Discussion

Numerous studies indicate that there is still no pipeline applicable to all situations in landslide susceptibility and hazard assessment (Chen et al. 2020; Suárez et al. 2020; Xie et al. 2021). In general, the improvement of the reliability and accuracy of the landslide hazard assessment results can be concentrated in two parts: better data and stronger model. As the data improvement is limited by the availability of data and the actual situation of the selected area, the improvement of assessment models becomes particularly crucial. In many studies, many algorithms are considered to compare their performance, while ignoring the improvement of a single model itself (Akgun 2011; Tien Bui et al. 2012; Erener et al. 2016) compared the GIS-based multi-criteria decision analysis, logistic regression (LR), and association rule mining, and the results showed that LR methods were better than other methods. Such a comparison strategy is certainly useful; however, it ignores the optimization of the model. The enhancement of hyperparameters is crucial for the optimization of the model. The performance of one machine learning model without the optimal hyperparameters significantly reduces compared to the model with the best hyperparameters. However, in the field of machine learning, according to the No Free Lunch principle, no perfect set of hyperparameters fits all the models (Snoek et al. 2012; Wang et al. 2019). Therefore, the tuning of hyperparameters is extremely important for the machine learning-based landslide hazard assessment.

For this purpose, in this study, a new model named BO-SVM based on the BO algorithm was proposed. In theory, replacing the empirical risk minimization principle in the traditional methods with the structural risk minimization principle, the BO algorithm obtains the overall optimal hyperparameters of the model through the Gaussian process to improve the performance of the model. In practice, in the case of the same input dataset, optimization of the hyperparameters of the model through the BO algorithm showed improved landslide hazard assessment result. The prediction effect of the BO-SVM model is higher than that of the SVM model, and the prediction accuracy and AUC value increased by 5 and 4 %, respectively.

In this study, the key step to run BO was to realize the Gaussian process regression algorithm and optimize the computational process through the kernel trick of SVM. In general, common optimization algorithms cannot make full use of all the known results, and the potential relationships between the known results may be ignored as well (Nhu et al. 2020). Figure 5 shows that the AUC value obtained by the BO is generally show a broader distribution compared to random search in the iterative process. At the beginning



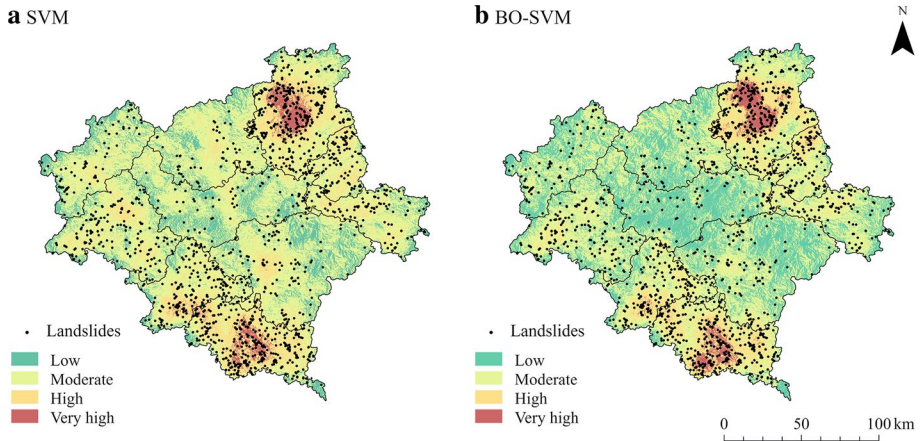
**Fig. 5** **a** The AUC value of BO and random search in iteration; **b** ROC curves of SVM and BO-SVM, AUC is the acronym of area under the ROC curve

**Table 3** Statistic result of two landslide hazard models

Hazard zone	Model	Grid number	Area proportion (%)	Landslide number	Landslide proportion (%)
Low	SVM	548,968	15.96	79	4.62
	BO-SVM	1,891,009	54.98	102	5.96
Moderate	SVM	2,193,576	63.78	675	39.45
	BO-SVM	1,038,155	30.19	499	29.16
High	SVM	602,680	17.52	698	40.79
	BO-SVM	410,526	11.94	702	41.03
Very high	SVM	93,946	2.73	259	15.14
	BO-SVM	99,480	2.89	408	23.85

of the iteration, the distribution of the AUC values for both the algorithms showed a random state. However, as the iterative process proceeds, the trend of BO gradually increases, while the random search gradually decreases. Eventually, the BO significantly outperforms random search, as it is a global optimization method that can be adjusted for the next optimization with the help of a Gaussian function, while random search does not have such an adjustment.

In the entire study area, the hazard zones’ spatial distribution of the two models is roughly similar (Fig. 6), because the difference between the models is not large fundamentally based on dualistic statistics. The very high hazard zone areas are distributed in the southwest and northeast of the study area (Yanping District and Pucheng County) and have suffered the most landslide disasters in history. The low hazard zones and moderate hazard zones are most widely distributed in the study area, reporting a few landslide disasters in these areas, but the results of the two models showed that they are far from reaching a high risk. On a local scale, the results of the two models have some noticeable differences. The moderate zones area of the SVM model is significantly



**Fig. 6** The landslide hazard map of the two models

larger than that of the BO-SVM and are usually called the uncertainty interval, indicating that the model has a low degree of confidence in the occurrence of the landslide (Sun et al. 2020). The result of SVM showed many central parts of the study area including most parts of Wuyishan County, Jianyang District, and Jianou County. However, there are some scars of high zone along the railways, with few landslides reported in these areas. The BO-SVM produced less uncertain intervals and maintained a high accuracy in the areas along the railways. The above results show that the BO-SVM outperforms the SVM, indicating improvement over the BO model. In this study, the landslide hazard maps obtained by the two models differed significantly at the local scale, and these differences occurred where BO enhanced the machine learning model. Currently, a few studies have reached similar conclusions. Yang et al. (2019b) proposed a hybrid model based on the Bayesian theory, exhibiting that this hybrid model performed better than the traditional models at local scales.

The optimization results of the approach used in this study are theoretically and practically superior, and similar findings have been made in other studies. Chen et al. (2017a) proposed a method to optimize landslide spatial modeling with genetic algorithm (GA), differential evolution (DE), and particle swarm optimization (PSO), indicating that the optimized models show some improvement relative to the original models. This further proves the feasibility and superiority of the approach used in this study. However, different optimization algorithms have different effects on different machine learning models. In theory, BO as a global optimization model is more effective than local optimization models such as PSO.

To some extent, this study still had some limitations. The entire model has a certain black-box nature, and the optimization process relies on the “trial and error” of the computer, rather than calculating towards a visible goal. As mentioned above, the hyperparameters are important aspect of the model, and optimization algorithms can solve for optimal hyperparameters. By analogy, the optimization algorithm itself also has hyperparameters, which might be called “hyper-hyperparameters”. The presence of these “hyper-hyperparameters” also affects the optimization algorithm and the target model. However, there is no additional way to compute these parameters. Perhaps reducing this laborious optimization process is worthy of attention. On the other hand, the research only uses random

search as a comparison of the BO algorithm, while ignoring methods such as grid search. The future study will focus on overcoming the deficiencies of this model.

## 6 Conclusions

In conclusion, an advanced approach based on the BO algorithm and SVM for landslide hazard assessment was successfully developed, aiming to use a hyperparameter optimization method (BO algorithm) to solve the problem of machine learning hyperparameter selection in the landslide hazard assessment. An investigation into BO-SVM and SVM as the assessment model for landslides hazards assessment of Nanping city indicated better performance of BO-SVM model than those of SVM. For the BO-SVM model, the accuracy of the confusion matrix and the AUC value of the ROC curve was 89.53% and 0.97, respectively. In the landslide hazard zoning map generated by the two models, the BO-SVM was also found more reliable; 65% of the historical landslides are in very high and high hazard areas, which together cover less than 15% of the study area. The BO-SVM significantly performed better than the SVM in the classification of landslide hazard at the local scale, proving that the BO algorithm has a significant optimization effect on the model. The findings of this study could provide a rational perspective for improving the landslide hazard assessment and is of certain significance to other landslide studies using machine learning methods. In addition, the results of this method could be helpful for risk assessment and management of other natural disasters.

**Acknowledgements** We are grateful to the anonymous reviewers and the Editor for their constructive comments that helped us improve the quality of the paper. We sincerely acknowledge the data support from the “National Earth System Science Data Center, National Science & Technology Infrastructure of China. (<http://www.geodata.cn>)”, International Scientific & Technical Data Mirror Site, Computer Network Information Center, Chinese Academy of Sciences. (<http://www.gscloud.cn>).

**Funding** The work was supported by the National Natural Science Foundation of China (No. 41861134011) and (No. 51874268).

**Data availability** All the data and material in this paper are available from the Internet and the URL where the data were obtained has been shown in the text.

**Code availability** Code Non-Public.

**Declarations**

**Conflict of interest** The authors declare no competing financial interests.

## References

- Abdollahi S, Pourghasemi HR, Ghanbarian GA et al (2018) Prioritization of effective factors in the occurrence of land subsidence and its susceptibility mapping using an SVM model and their different kernel functions[J]. *Bull Eng Geol Environ* 78(6):4017–4034
- Akgun A (2011) A comparison of landslide susceptibility maps produced by logistic regression, multi-criteria decision, and likelihood ratio methods: a case study at İzmir, Turkey[J] *Landslides* 9(1):93–106

- Ambrosi C, Strozzi T, Scapozza C et al (2018) Landslide hazard assessment in the Himalayas (Nepal and Bhutan) based on Earth-Observation data[J]. *Eng Geol* 237(1):217–228
- An K, Kim S, Chae T et al (2018) Developing an accessible landslide susceptibility model using open-source resources[J]. *Sustainability* 10(2):293
- Ballabio C, Sterlacchini S (2012) Support Vector Machines for Landslide Susceptibility Mapping: The Staffora River Basin Case Study, Italy[J]. *Math Geosci* 44(1):47–70
- Bourenane H, Guettouche MS, Bouhadad Y et al (2016) Landslide hazard mapping in the Constantine city, Northeast Algeria using frequency ratio, weighting factor, logistic regression, weights of evidence, and analytical hierarchy process methods[J]. *Arab J Geosci* 9(2):24
- Chang CC, Lin CJ (2011) LIBSVM: A Library for Support Vector Machines[J]. *ACM Trans Intell Syst Technol* 2(3):1–27
- Chen W, Chen X, Peng JB et al (2021a) Landslide susceptibility modeling based on ANFIS with teaching-learning-based optimization and Satin bowerbird optimizer[J]. *Geosci Front* 12(1):93–107
- Chen W, Chen YZ, Tsangaratos P et al (2020) Combining evolutionary algorithms and machine learning models in landslide susceptibility assessments[J]. *Remote Sens* 12(23):3854
- Chen W, Lei X, Chakraborty R et al (2021b) Evaluation of different boosting ensemble machine learning models and novel deep learning and boosting framework for head-cut gully erosion susceptibility[J]. *J Environ Manage* 284:112015
- Chen W, Li Y (2020) GIS-based evaluation of landslide susceptibility using hybrid computational intelligence models[J]. *Catena* 195:104777
- Chen W, Panahi M, Pourghasemi HR (2017a) Performance evaluation of GIS-based new ensemble data mining techniques of adaptive neuro-fuzzy inference system (ANFIS) with genetic algorithm (GA), differential evolution (DE), and particle swarm optimization (PSO) for landslide spatial modelling[J]. *Catena* 157:310–324
- Chen W, Pourghasemi HR, Kornejady A et al (2017b) Landslide spatial modeling: Introducing new ensembles of ANN, MaxEnt, and SVM machine learning techniques[J]. *Geoderma* 305:314–327
- Chen YZ, Chen W, Janizadeh S et al (2021c) Deep learning and boosting framework for piping erosion susceptibility modeling: spatial evaluation of agricultural areas in the semi-arid region[J]. *Geocarto Int*: 1–27
- Cortes C, Vapnik V (1995) Support-Vector Networks[J]. *Mach Learn* 20(3):273–297
- Den Eeckhaut MV, Marre A, Poesen J (2010) Comparison of two landslide susceptibility assessments in the Champagne–Ardenne region (France)[J]. *Geomorphology* 115(1–2):141–155
- Dou J, Yunus AP, Bui DT et al (2020) Improved landslide assessment using support vector machine with bagging, boosting, and stacking ensemble machine learning framework in a mountainous watershed. *Japan[J] Landslides* 17(3):641–658
- Erener A, Mutlu A, Duzgun HS (2016) A comparative study for landslide susceptibility mapping using GIS-based multi-criteria decision analysis (MCDA), logistic regression (LR) and association rule mining (ARM)[J]. *Eng Geol* 203:45–55
- Fan XM, Scaringi G, Korup O et al (2019) Earthquake-Induced Chains of Geologic Hazards: Patterns, Mechanisms, and Impacts[J]. *Rev Geophys* 57(2):421–503
- Frazier PI (2018) A tutorial on bayesian optimization[J]. *arXiv preprint arXiv:180702811*
- Froude MJ, Petley DN (2018) Global fatal landslide occurrence from 2004 to 2016[J]. *Nat Hazards Earth Syst Sci* 18(8):2161–2181
- Gariano SL, Guzzetti F (2016) Landslides in a changing climate[J]. *Earth-Sci Rev* 162:227–252
- Greenhill S, Rana S, Gupta S et al (2020) Bayesian Optimization for Adaptive Experimental Design: A Review[J]. *IEEE Access* 8:13937–13948
- He J, Qiu H, Qu F et al (2021) Prediction of spatiotemporal stability and rainfall threshold of shallow landslides using the TRIGRS and Scoops3D models[J]. *Catena* 197:104999
- Hung O, Leroueil S, Picarelli L (2013) The Varnes classification of landslide types, an update[J]. *Landslides* 11(2):167–194
- Intrieri E, Carla T, Gigli G (2019) Forecasting the time of failure of landslides at slope-scale: A literature review[J]. *Earth-Sci Rev* 193:333–349
- Jafarian Y, Lashgari A, Haddad A (2019) Predictive Model and Probabilistic Assessment of Sliding Displacement for Regional Scale Seismic Landslide Hazard Estimation in Iran[J]. *Bull Seismol Soc Amer* 109(5):1581–1593
- Kirschbaum DB, Adler R, Hong Y et al (2009) A global landslide catalog for hazard applications: method, results, and limitations[J]. *Nat Hazards* 52(3):561–575
- Kobliha M, Schwarz J, Ocenasek J (2006) Bayesian optimization algorithms for dynamic problems. In: Rothlauf F (ed) *Applications of Evolutionary Computing, Proceedings*, vol 3907. *Lecture Notes in Computer Science*. 800–804



- Lee JH, Sameen MI, Pradhan B et al (2018) Modeling landslide susceptibility in data-scarce environments using optimized data mining and statistical methods[J]. *Geomorphology* 303:284–298
- Lo CM, Feng ZY, Chang KT (2018) Landslide hazard zoning based on numerical simulation and hazard assessment[J]. *Geomat Nat Hazards Risk* 9(1):368–388
- Luo X, Lin F, Zhu S et al (2019) Mine landslide susceptibility assessment using IVM, ANN and SVM models considering the contribution of affecting factors[J]. *PLoS One* 14(4):e0215134
- Marjanović M, Kovačević M, Bajat B et al (2011) Landslide susceptibility assessment using SVM machine learning algorithm[J]. *Eng Geol* 123(3):225–234
- Merghadi A, Yunus AP, Dou J et al (2020) Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance[J]. *Earth-Sci Rev* 2020:103225
- Moresi FV, Maesano M, Collalti A et al (2020) Mapping Landslide Prediction through a GIS-Based Model: A Case Study in a Catchment in Southern Italy[J]. *Geosciences* 10(8):309
- Nhu VH, Hoang ND, Nguyen H et al (2020) Effectiveness assessment of Keras based deep learning with different robust optimization algorithms for shallow landslide susceptibility mapping at tropical area[J]. *Catena* 188:104458
- Olen S, Bookhagen B (2018) Mapping Damage-Affected Areas after Natural Hazard Events Using Sentinel-1 Coherence Time Series[J]. *Remote Sens* 10(8):19
- Pal SC, Chowdhuri I (2019) GIS-based spatial prediction of landslide susceptibility using frequency ratio model of Lachung River basin, North Sikkim, India[J]. *SN Applied Sciences* 1(5):416
- Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: Machine Learning in Python[J]. *J Machine Learn Res* 12:2825–2830
- Phong TV, Phan TT, Prakash I et al (2019) Landslide susceptibility modeling using different artificial intelligence methods: a case study at Muong Lay district, Vietnam[J]. *Geocarto Int*: 1–24
- Qiu HJ, Cui YF, Yang DD et al (2019) Spatiotemporal Distribution of Nonseismic Landslides during the Last 22 Years in Shaanxi Province, China[J]. *ISPRS Int Geo-Inf* 8(11):20
- Reichenbach P, Rossi M, Malamud BD et al (2018) A review of statistically-based landslide susceptibility models[J]. *Earth-Sci Rev* 180:60–91
- Smola AJ, Scholkopf B (2004) A tutorial on support vector regression[J]. *Stat Comput* 14(3):199–222
- Snoek J, Larochelle H, Adams RP (2012) Practical bayesian optimization of machine learning algorithms. In: *Adv neural Inf Process Syst*. 2951–2959
- Sorbino G, Sica C, Cascini L (2009) Susceptibility analysis of shallow landslides source areas using physically based models[J]. *Nat Hazards* 53(2):313–332
- Suárez G, Domínguez-Cuesta MJ (2020) Improving landslide susceptibility predictive power through colluvium mapping in Tegucigalpa, Honduras[J]. *Nat Hazards* 105(1):47–66
- Sun DL, Wen HJ, Wang DZ et al (2020) A random forest model of landslide susceptibility mapping based on hyperparameter optimization using Bayes algorithm[J]. *Geomorphology* 362(2020):107201
- Tang XZ, Hong HY, Shu YQ et al (2019) Urban waterlogging susceptibility assessment based on a PSO-SVM method using a novel repeatedly random sampling idea to select negative samples[J]. *J Hydrol* 576:583–595
- Tehrany MS, Pradhan B, Mensor S et al (2015) Flood susceptibility assessment using GIS-based support vector machine model with different kernel types[J]. *Catena* 125:91–101
- Theron A, Engelbrecht J (2018) The Role of Earth Observation, with a Focus on SAR Interferometry, for Sinkhole Hazard Assessment[J]. *Remote Sens* 10(10):30
- Tien Bui D, Pradhan B, Lofman O et al (2012) Landslide susceptibility assessment in the Hoa Binh province of Vietnam: A comparison of the Levenberg–Marquardt and Bayesian regularized neural networks[J]. *Geomorphology* 171:12–29
- Tien Bui D, Shahabi H, Shirzadi A et al (2018) Landslide Detection and Susceptibility Mapping by AIR-SAR Data Using Support Vector Machine and Index of Entropy Models in Cameron Highlands, Malaysia[J]. *Remote Sens* 10(10):32
- van Westen CJ, Castellanos E, Kuriakose SL (2008) Spatial data for landslide susceptibility, hazard, and vulnerability assessment: An overview[J]. *Eng Geol* 102(3–4):112–131
- Violante RA, Bozzano G, Rovere EI (2018) The Marine Environment: Hazards, Resources and the Application of Geoethics Principles[J]. *Ann Geophys* 60:1–10
- Wang D, Wang CC, Xiao JH et al (2019) Bayesian optimization of support vector machine for regression prediction of short-term traffic flow[J]. *Intell Data Analy* 23(2):481–497
- Wu D, Huang MX, Zhang Y et al (2018) Strategy for assessment of disaster risk using typhoon hazards modeling based on chlorophyll-a content of seawater[J]. *EURASIP J Wirel Commun Netw* 2018(1):12

- Xie W, Li XS, Jian WB et al (2021) A Novel Hybrid Method for Landslide Susceptibility Mapping-Based GeoDetector and Machine Learning Cluster: A Case of Xiaojin County, China[J]. *ISPRS Int Geo-Inf* 10(2):93
- Xu C, Xu XW, Lee YH et al (2012) The 2010 Yushu earthquake triggered landslide hazard mapping using GIS and weight of evidence modeling[J]. *Environ Earth Sci* 66(6):1603–1616
- Yanar T, Kocaman S, Gokceoglu C (2020) Use of Mamdani Fuzzy Algorithm for Multi-Hazard Susceptibility Assessment in a Developing Urban Settlement (Mamak, Ankara, Turkey)[J]. *ISPRS Int Geo-Inf* 9(2):114–139
- Yang JT, Song C, Yang Y et al (2019a) New method for landslide susceptibility mapping supported by spatial logistic regression and GeoDetector: A case study of Duwen Highway Basin, Sichuan Province, China[J]. *Geomorphology* 324:62–71
- Yang Y, Yang JT, Xu CD et al (2019b) Local-scale landslide susceptibility mapping using the B-GeoSVC model[J]. *Landslides* 16(7):1301–1312
- Yin J, Yin Z, Xu SY (2013) Composite risk assessment of typhoon-induced disaster for China's coastal area[J]. *Nat Hazards* 69(3):1423–1434
- Youssef AM, Pourghasemi HR, Pourtaghi ZS et al (2015) Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia[J]. *Landslides* 13(5):839–856
- Zezere JL, Pereira S, Melo R et al (2017) Mapping landslide susceptibility using data-driven methods[J]. *Sci Total Environ* 589:250–267
- Zhang FY, Huang XW (2018) Trend and spatiotemporal distribution of fatal landslides triggered by non-seismic effects in China[J]. *Landslides* 15(8):1663–1674
- Zhao X, Chen W (2020) Optimization of Computational Intelligence Models for Landslide Susceptibility Evaluation[J]. *Remote Sens* 12(14):2180

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.